All The Cool Kids, How Do They Fit In? Popularity and Demographic Biases in Recommender Evaluation and Effectiveness[*][†]

Michael D. Ekstrand MICHAELEKSTRAND@BOISESTATE.EDU
Mucun Tian MUCUNTIAN@U.BOISESTATE.EDU
Ion Madrazo Azpiazu IONMADRAZO@U.BOISESTATE.EDU
Jennifer D. Ekstrand JENNIFEREKSTRAND@U.BOISESTATE.EDU
Oghenemaro Anuyah OGHENEMAROANUYAH@U.BOISESTATE.EDU
David McNeill DAVIDMCNEILL@U.BOISESTATE.EDU
Maria Soledad Pera SOLEPERA@BOISESTATE.EDU
People and Information Research Team, Dept. of Computer Science, Boise State University

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

In the research literature, evaluations of recommender system effectiveness typically report results over a given data set, providing an aggregate measure of effectiveness over each instance (e.g. user) in the data set. Recent advances in information retrieval evaluation, however, demonstrate the importance of considering the distribution of effectiveness across diverse groups of varying sizes. For example, do users of different ages or genders obtain similar utility from the system, particularly if their group is a relatively small subset of the user base? We apply this consideration to recommender systems, using offline evaluation and a utility-based metric of recommendation effectiveness to explore whether different user demographic groups experience similar recommendation accuracy. We find demographic differences in measured recommender effectiveness across two data sets containing different types of feedback in different domains; these differences sometimes, but not always, correlate with the size of the user group in question. Demographic effects also have a complex—and likely detrimental—interaction with popularity bias, a known deficiency of recommender evaluation. These results demonstrate the need for recommender system evaluation protocols that explicitly quantify the degree to which the system is meeting the information needs of all its users, as well as the need for researchers and operators to move beyond naïve evaluations that favor the needs of larger subsets of the user population while ignoring smaller subsets.

Keywords: recommender systems, fair evaluation

1. Introduction

Recommender systems are algorithmic tools for identifying items (e.g., products or services) of interest to users (Adomavicius and Tuzhilin, 2005; Ekstrand et al., 2010; Ricci et al., 2015). They are usually deployed to help mitigate information overload (Resnick et al., 1994). Internet-scale item spaces offer many more choices than humans can process, diminishing the quality of their decision-making abilities (Toffler, 1990; Gross, 1964). Recommender systems alleviate this problem by allowing users to more quickly focus on items likely to match their particular tastes. They are deployed across the modern Internet, suggesting products in e-commerce sites, movies and music in streaming media platforms, new connections on social networks, and many more types of items.

We are concerned with the *fairness* of recommender systems, a surprisingly tricky concept to define. In addition to the numerous types and op-

[*] This paper can be reproduced with scripts available at https://dx.doi.org/10.18122/B2GM6F.

[†] This paper is an extension of the poster by Ekstrand and Pera (2017).

erationalizations of fairness in the research literature, recommender fairness must identify which stakeholder groups to consider for fair treatment (Burke, 2017).

Both offline (Herlocker et al., 2004; Shani and Gunawardana, 2011) and online (Knijnenburg et al., 2012) evaluations of recommender systems typically focus on evaluating the system's effectiveness in aggregate over the entire population of users. While individual user characteristics are sometimes taken into account, as in demographic-informed recommendation (Pazzani, 1999; Ghazanfar and Prugel-Bennett, 2010), the end evaluation still aggregates over all users.

Recent developments in human-centered information retrieval have incorporated user demographics and characteristics to evaluate search engines and understand users' search behavior. Weber and Castillo (2010) use light user information augmented with census-based demographics to understand who is using a search engine. Mehrotra et al. (2017) follow this trend by measuring Bing's ability to satisfy the information needs of different subgroups of its user population, e.g. assessing whether it meets the needs of grandparents as effectively as those of young professionals.

This attention is necessary because the largest subgroup of users will tend to dominate overall statistics. If other subgroups have different needs, their satisfaction will carry less weight in the final analysis. This can lead to a misguided perception of the performance of the system and, more importantly, make it more difficult to identify how to better serve specific demographic groups.

Our fundamental research question is this: *Do different demographic groups obtain different utility from the recommender system?* This is a starting point for many further questions, such as whether particular demographic groups need to be better served by recommender systems and, if so, how they can be identified and supported in their information needs.

To address this question, we present an empirical analysis of the effectiveness of collaborative filtering recommendation strategies, stratified by the gender and age of the users in the data set. We apply widely-used recommendation techniques across two domains, musical artists and movies, using publicly-available data. We also explore the effect of rebalancing the data set by gender, the influence of user profile size on recommendation quality, and the interaction of demographic effects with previously documented biases in recommender evaluation, all in the context of demographically-distributed differences in effectiveness.

Our work is inspired by that of Mehrotra et al. (2017). We translate the concepts of their analysis from search engines to recommender systems. While our experiment is less sophisticated than Mehrotra et al.'s and necessarily limited by our offline experimental setting, it is fully reproducible using widely-distributed public data sets and can be easily adapted to additional algorithms, domains, and applications.

2. Background and Related Work

Recommender systems (Adomavicius and Tuzhilin, 2005; Ekstrand et al., 2010) are algorithmic tools for helping users find items that they may wish to purchase or consume. They have substantial influence; the best available public data indicates that recommendation drives 85% of Netflix video viewing (Gomez-Uribe and Hunt, 2015) and 30% of Amazon purchases (Linden et al., 2003).

2.1. Recommendation Techniques

There are a variety of families of recommendation algorithms. *Collaborative filters* (Ekstrand et al., 2010) mine user-item interaction traces, such as purchase records, click logs, or user-provided ratings of items, to generate recommendations based on the behavior of other users with similar taste. *Content-based filters* (Pazzani and Billsus, 2007; Lops et al., 2011) use item content or metadata, such as tags and text, to recommend items with similar content to items the user has liked in the past. Many production systems use a combination of these and other techniques as *hybrid* strategies to enhance the overall recommendation process (Burke, 2002; Bobadilla et al., 2013).

2.2. Recommender System Evaluation

Recommender systems are evaluated in offline settings using evaluation protocols derived from information retrieval (Herlocker et al., 2004; Gunawardana and Shani, 2009; Bellogin, 2012). These protocols hide a portion of the data and attempt to predict it using the recommendation model, measuring either the model's ability to predict withheld ratings (*prediction accuracy* evaluation) or its ability to recommend withheld items (*top-N* evaluation).

Top-N evaluation is widely regarded as the preferred setting, as it reflects the end goal of the recommender system—to recommend items the user will like—more accurately than predicting ratings. Offline top-N evaluation, however, has significant known problems. Among these are *popularity bias* (Bellogin et al., 2011), where the evaluation protocol gives higher accuracy scores to algorithms that favor popular items irrespective of their ability to meet user information needs, and *misclassified decoys* (Ekstrand and Mahant, 2017; Cremonesi et al., 2010), where a good recommendation is erroneously penalized because data on user preferences is incomplete.

Online evaluation, commonly using A/B tests (Kohavi et al., 2007) and measuring user response to recommendation, is the gold standard for effectiveness and avoids many of the problems of offline evaluation. User studies (Knijnenburg et al., 2012) allow even deeper insight into *why* users respond to recommendations in the way that they do. This type of study, however, is more expensive to conduct (in terms of time, protocols, and resources) than its offline counterpart (Shani and Gunawardana, 2011).

2.3. Fairness in Recommender Systems

The recommender system research community has long been interested in examining the social dimension of recommendation; the earliest modern recommender systems were developed in a human-computer interaction setting (Resnick et al., 1994; Hill et al., 1995), and there has been work on how they promote diversity or balkanization (van Alstyne and Brynjolfsson, 2005; Hosanagar et al., 2013).

More recent work has begun to consider questions of fairness in recommendation. Proposals for fair recommendation methods include penalizing algorithms for disparate distribution of prediction error (Yao and Huang, 2017), balancing neighborhoods before producing recommendations (Burke et al., 2017), and making recommended items independent from protected information (Kamishima and Akaho, 2017).

Burke (2017) taxonomizes fairness objectives and methods based on which set of stakeholders in the recommender system are being considered, as it is meaningful to consider fairness among many groups in a recommender system. In our work, we examine the *C-fairness* of recommender algorithms: whether or not they treat their users (consumers) fairly.

2.4. Demographic-Aware Evaluation

Traditionally, demographic information have been considered in the past to improve the effectiveness of diverse tasks, from text classification (Hovy, 2015), to search (Weber and Castillo, 2010), and recommendation (Said et al., 2011). Unfortunately, little is known about the effects of demographic information when it comes to evaluation tasks (Langer and Beel, 2014).

Typical evaluations average over all users or data points, providing a simple aggregate measurement of the recommender's effectiveness. However, user satisfaction in a recommender system depends on more than accuracy (Herlocker et al., 2004; Langer and Beel, 2014). In fact, Mehrotra et al. (2017) demonstrate that this naïve approach to simply aggregate measurements masks important differences in how different groups of users experience the system. The system may be delivering high-quality service to one subset of its user group, while another smaller group of users receives lower-quality recommendations or search results; the overall metric will not reward effort that improves the experience of minority users as much as it rewards efforts that make things better for those already well-served.

The fundamental thrust of our present work is to translate this idea from the online web search setting employed by Mehrotra et al. to offline evaluation of recommender systems, and examine whether applying existing algorithms to existing public data sets will provide comparable utility to different groups of users. The discussion presented in this paper expands the initial analysis presented by Ekstrand and Pera (2017).

3. Data and Methods

We used the LensKit recommender toolkit (Ekstrand et al., 2011) to build and evaluate several collaborative filtering algorithms across multiple public data sets with different types of feedback in multiple product domains.

3.1. Data Sets

While there are many public records of ratings, plays, and other common recommender inputs for use in research, few of them have the necessary user demographic information to assess bias in recommender effectiveness. We have found three that have the necessary data: early versions of the MovieLens data (Harper and Konstan, 2016) and the two Last.FM data sets collected by Celma (2010). Table 1 summarizes these data sets.

Table 1: Summary of data sets

Datasets	Users	Items	Pairs	Density
LFM1K	992	177,023	904,625	0.52%
LFM360K	359,347	160,168	17,559,443	0.03%
ML1M	6,040	3,706	1,000,209	4.47%

The LFM1K data set contains 19M records of 992 users playing songs from 177K artists, gathered from the Last.FM audioscrobbler. We aggregated this data at the artist level to produce play counts for 904K user-artist pairs. The LFM360K data set contains the top 50 most-played artists from 360K users along with their play counts, covering 160K artists. Both data sets contain gender, age, and sign-up date for many users (Figure 1 shows demographic coverage).

The ML1M data set contains 1M 5-star ratings of 3,900 movies by 6,040 users who joined MovieLens, a noncommercial movie recommendation service operated by the University of Minnesota, through the year 2000. Each user has a self-reported age, gender, occupation, and zip code. Some time after releasing the 1M data set, MovieLens stopped collecting demographic data from new users, so the larger recent data sets (10M and 20M) do not contain the data required for our experiment.

3.2. Source Data Distributions

Differences in recommender effectiveness need to be understood in the context of the demographic distribution of the underlying data.

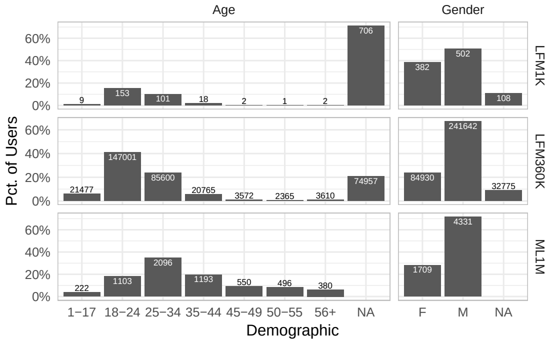


Figure 1: User distribution by demographic group. Numbers in bars are the number of users in that bin.

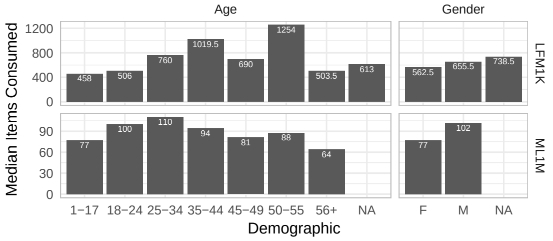


Figure 2: Median items consumed by users in each demographic group. We omit LFM360K since it only contains each user's top 50 artists.

Figure 1 shows the distribution of each data set. All three data sets exhibit similar distributions of user genders, with the majority of users reporting as male; LFM1K is the least imbalanced. The largest block of ML1M users belong to the [25-35] group, whereas a plurality of LFM360K users belong to the [18-24] group; most LFM1K users did not report their age. Approximately 10% of LFM360K users declined to share their gender while close to 20% declined to share their age. All user records in the ML1M data set contain full demographic information. For consistency among the reported results, we bin Last.FM users into the same age groups used in the ML1M data set throughout.

Figure 2 shows user activity levels, as measured by the number of movies rated or artists played, in each user group. Men are more active than women in both data sets. The activity-age relationship in ML1M data almost follows the demographic distribution, with those groups that have more users also having more active users; the small number of users in most age brackets in LFM1K preclude drawing conclusions from age-activity relationships in that data.

3.3. Experimental Protocol

We partitioned each data set with 5-fold cross-validation. Our primary results use LensKit's default user-based sampling strategy: select 5 test sets of users, and for each user select 5 ratings to be the test ratings; the rest of those users' ratings, along with all ratings from users not in that test set, comprise the train set for that test set. For LFM360K, we sampled 5 disjoint sets of 5000 test users (or items) for each test set to decrease compute time. For LFM1K and ML1M, we partitioned the users into 5 disjoint sets.

We also tested Bellogin's U1R method (Bellogin, 2012) for neutralizing popularity bias; this works exactly like the default, except it picks test sets of items instead of users, and it generates a different recommendation list for each user-item pair in the test data, with that item as the only test item to be found. The idea is that, by having the same number of test ratings for each item, recommenders that favor popular items can't win simply by having popular items be the right answer more often than unpopular ones.

3.4. Performance Metrics

We measure recommender effectiveness using Normalized Discounted Cumulative Gain (nDCG) (Järvelin and Kekäläinen, 2002), a widely-accepted measure of the effectiveness of a recommender system. nDCG measures the utility that a user is expected to obtain from a recommendation list, based on that user's estimated utility for individual items and the position in the list at which those items were presented. The nDCG for a recommendation list L generated for user u is computed with Equation 1:

$$\mathrm{nDCG}_{L,u} = \frac{\mathrm{DCG}_{L,u}}{\mathrm{IDCG}_u} \tag{1}$$

$\mathrm{DCG}_{L,u}$ is defined by Equation 2, where l_i is the i-th item in list L and $\mu_u(l_i)$ is user u's utility for item l_i, and IDCG_u is computed as $\mathrm{DCG}_{L,u}$, with a list consisting only of the user's rated items in non-increasing order of utility.

$$\mathrm{DCG}_{L,u} = \mu_u(l_1) + \sum_{i=2}^{|L|} \frac{\mu_u(l_i)}{\log_2 i} \tag{2}$$

$\mathrm{nDCG}_{L,u}$ quantifies the utility achieved by a recommendation list as a fraction of the total achievable utility if the recommender could perfectly identify the user's most-preferred items. For the ML1M data set, we define $\mu_u(l_i)$ as the user's rating for movie l_i; for the Last.FM data sets, we use the number of times the user has played the artist. Items for which no data is available are assumed to have a utility of 0. Although this has significant conceptual problems (Ekstrand and Mahant, 2017), it is standard practice in recommender systems research, and there is no widely-accepted improvement.

3.5. Algorithms

We employed several classical and widely-used collaborative filtering algorithms, as implemented by LensKit. We operated each algorithm in both explicit (rating-based) and implicit (consumption record) feedback mode.

- *Popular* (Pop), recommending the most frequently-consumed items.
- *Mean*, recommending the items with the highest average rating.
- *Item-Item* (II), an item-based collaborative filter (Sarwar et al., 2001; Deshpande and Karypis, 2004) using 20 neighbors and cosine similarity. The explicit-feedback version normalizes ratings by subtracting item means; the implicit-feedback version replaces the weighted average with a simple sum of similarities.
- *User-User* (UU), a user-based collaborative filter (Resnick et al., 1994; Herlocker et al., 2002) configured to use 30 neighbors and cosine similarity. The explicit-feedback variant uses user-mean normalization for user rating vectors, and the implicit-feedback variant

again replaces weighted averages with sums of similarities. User-user did not provide effective recommendations on the Last.FM data, so we exclude it from that data set's results.

- *FunkSVD* (MF), the popular gradient descent matrix factorization technique (Funk, 2006; Paterek, 2007) with 40 latent features and 150 training iterations per feature.

In the results, each algorithm is tagged with its variant. Algorithms suffixed with '-E' are explicit-feedback recommenders (applicable only to ML); '-B' are implicit-feedback recommenders that only consider *whether* an item was rated or played, irrespective of the number of plays (both data sets); and '-C' are implicit-feedback recommenders that use the number of times an artist was played as repeated feedback (LFM1K and LFM360K), log-normalized prior to recommendation.

The purpose of this work is not to compare algorithms, but to compare recommender performance across demographic groups. We have selected these algorithms to provide a representative sample of classical collaborative filtering approaches.

4. Results

Using the data and methods presented in Section 3, we discuss below the results of the experiments conducted to quantify user satisfaction with presented recommendations among different demographic groups. For doing so, we consider three different perspectives that guide our assessments: (i) analysis based on raw data, i.e., considering all users in the data sets, (ii) analysis based on user activity levels, i.e., controlled profile size, and (iii) analysis based on gender-balanced data sets.

4.1. Basic Results

In order to quantify to what extent demographics affect the overall satisfaction obtained by the users, we conducted an experiment that considers the performance of traditional recommendation algorithms for different gender and age groups, respectively.

Figure 3 illustrates the overall satisfaction obtained by each gender group, measured by nDCG, whereas Figure 4 does the same for users grouped by age.

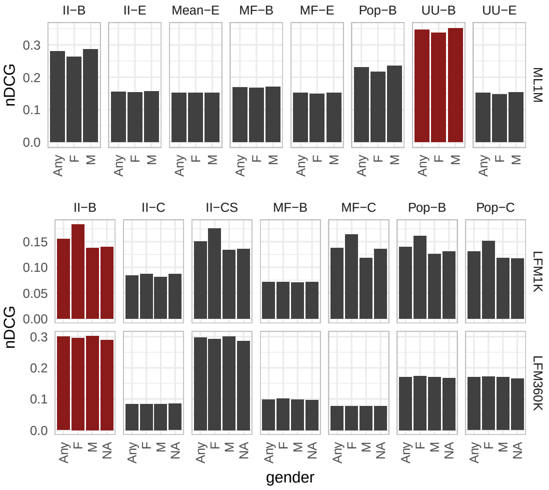


Figure 3: Algorithm performance by gender. Highlighted cell is for the algorithm with the best overall performance on that data set.

For each data set's best-performing algorithm (highlighted), we compared the differences in utility for each demographic group. ML1M and LFM1K have statistically-significant differences between gender groups, and LFM360K has significant differences between age brackets (Kruskal-Wallis $p < 0.01$ with the Bonferroni correction for multiple comparisons).

4.2. Controlling for Profile Size

As seen in Figure 2, different demographic groups have different activity levels as measured by the number of items they have rated or consumed. The size of a user's profile can be a factor in their recommendation utility, given that more items provide a stronger basis for recommendation. To control for the effect of profile size on user satisfaction, we fitted linear models predicting the nDCG using the number of items in the user's profile (excluding LFM360K, since it only contains each user's top 50 artists). We used the average nDCG achieved by all algorithms for a particular user as the dependent variable, so we are only predicting a single metric per user; this captures an overall notion of the 'difficulty' of pro-

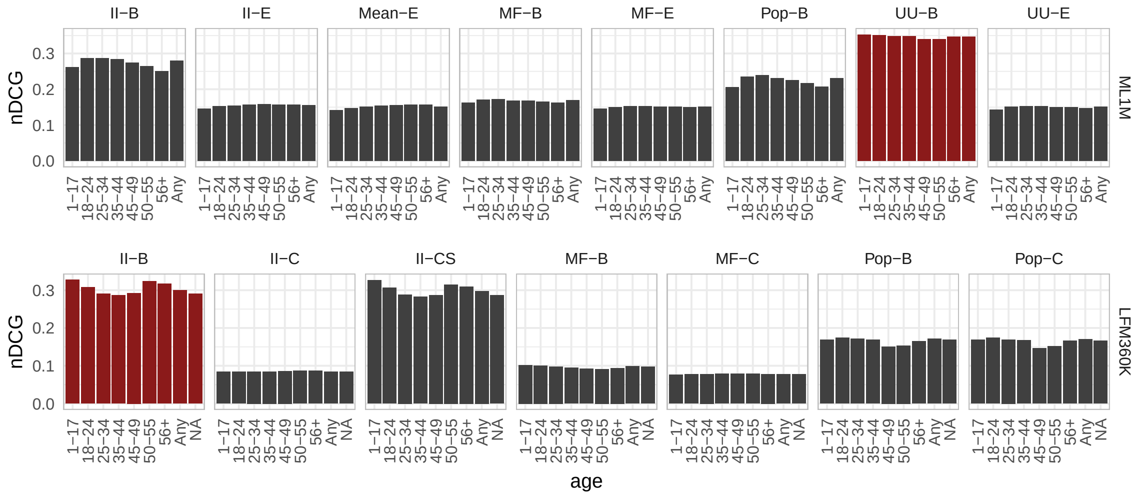


Figure 4: Algorithm performance by age. Highlighted cell has highest overall accuracy. We omit LFM1K because most users in that data set lack age data.

ducing effective recommendations for that user. Figure 5 shows the fitted models; we apply a log transform to the item count and take the square root of the nDCG to achieve a better fit. Surprisingly, there is a negative relationship between user profile size and recommendation accuracy; the exact cause is unknown, but we suspect that users with more items in their profile have already rated the 'easy' items, so recommending for them is a harder problem.

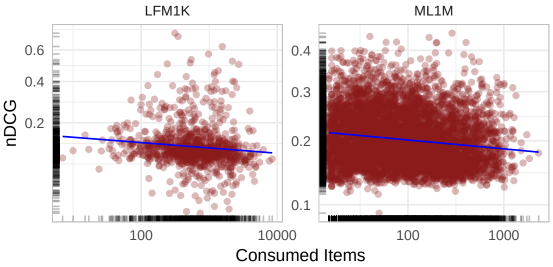
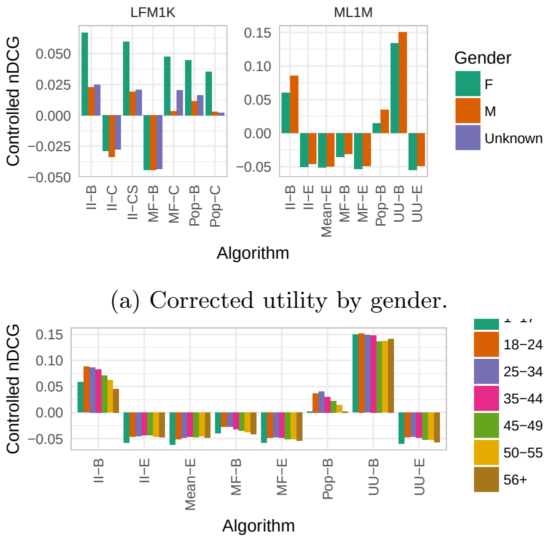


Figure 5: Models predicting nDCG with profile size.

Figure 6 shows the nDCG for each group after removing the effect of user profile size. We see that the demographic effects observed in Section 4.1 remain after this control, indicating a demographic effect of training the models on the data beyond that explained by user profile size.



(a) Corrected utility by gender.

(b) MovieLens corrected utility by age

Figure 6: Recommendation utility after controlling for profile size.

4.3. Resampling for Balance

As shown in Figure 1, both ML1M and LFM360K data sets include a larger proportion of male users, unbalancing the training data. As preprocessing data to produce fair training data is one way to train fair models (Kamiran and Calders, 2009), we resampled the ML1M and LFM360K data sets to produce gender-balanced versions of each and re-trained the algorithms.

We balanced the data sets by identifying users with known gender information, and randomly sampling without replacement the same number of female and male users (1500 samples each for the ML1M data set and 75000 samples each for LFM360K data set).

Figure 7 shows the experiment results on the gender-balanced data sets, and Table 2 shows the numeric change from the unbalanced experiment for the best-performing algorithm on each data set. We repeated the Kruskal-Wallis test on both sampled ML1M and LFM360K data sets, and it did not find a statistically significant difference between groups on either data set. Resampling the data, while reducing recommender accuracy slightly, did not create new gender differences in performance for LFM360K, and seems to have reduced the difference for ML1M. We are not sure that it went entirely away, as the Kruskal-Wallis test may be overly conservative and does not test directly for the elimination of an effect, but it does seem to have diminished. Resampling so that each group has the same number of ratings may eliminate the difference.

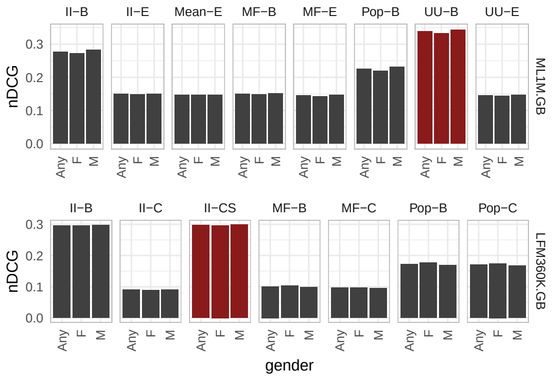


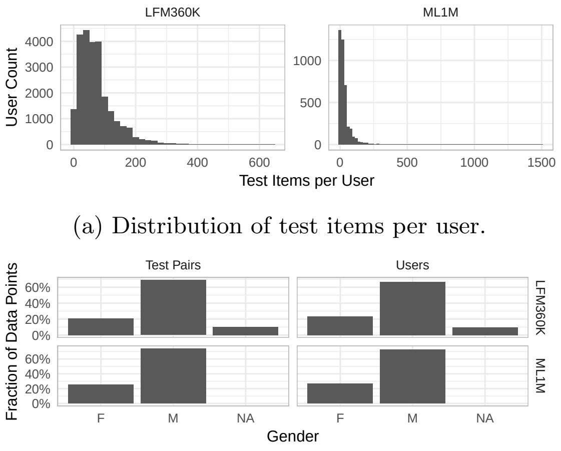
Figure 7: Algorithm accuracy by gender on balanced data sets. Highlighted cell is for the algorithm with the best overall performance on that data set.

Table 2: Changes on nDCG observed on balanced vs. raw data on ML1M and LFM360K data sets

Datasets	Algorithm	Gender	nDCG	nDCG (Balanced data)	Relative Difference
ML1M	UU-B	Female	0.337	0.334	1.03%
ML1M	UU-B	Male	0.351	0.344	2.22%
ML1M	UU-B	Any	0.347	0.339	2.54%
LFM360K	II-CS	Female	0.293	0.296	1.11%
LFM360K	II-CS	Male	0.301	0.298	0.76%
LFM360K	II-CS	Any	0.297	0.297	0.06%

4.4. Reducing Popularity Bias

To reduce the effect of popularity bias, we ran the ML1M version of the experiment using Bellogin's U1R protocol, as described in Section 3.3. Since this protocol partitions items instead of users, different users may have different numbers of test items, and the distribution of user demographics may differ from the underlying data. Figure 8a shows the distribution of test pairs per user, and Figure 8b shows the demographic distribution of the users in the test data. This distribution corresponds well to the underlying user distribution.



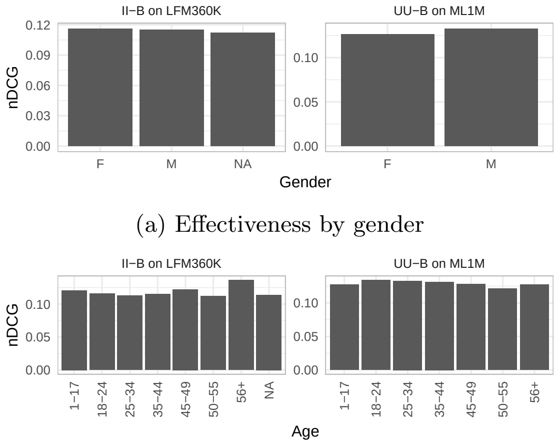
(a) Distribution of test items per user.

(b) Gender distribution of test pairs and test users.

Figure 8: Distribution of test data in ML1M U1R experiment.

Figure 9 shows accuracy by demographic group for the best algorithm for each data set under the U1R protocol. The differences on gender are consistent with the basic results in Figure 3. We compared two averaging strategies, averaging across all user-item pairs by user gender and averaging each user's recommendation results prior to averaging all users with a particular gender, and saw no difference.

Age tells a different story — on the LFM360K, we see a different pattern in the distribution of accuracy across ages than we do under the user-based evaluation protocol in Figure 4. It is not clear which provides a more accurate picture, but this does demonstrate that correcting for one effect (popularity bias) can change the results for another effect (demographic bias).



(a) Effectiveness by gender

(b) Effectiveness by age

Figure 9: Recommender effectiveness under U1R protocol. The best overall algorithm for each data set is shown.

5. Discussion and Limitations

Having observed some differences in recommender performance between demographic groups, we now turn to the implications of our results and some of their limitations.

5.1. Implications for Recommender Evaluation

The existence of differences in measured recommender performance between demographic groups indicates a need to consider *who* is obtaining *how much* benefit from a recommender system. If some users are underserved by the recommender, it may be indicative of an area for improvement, particularly if that group of users represents a market segment in which the recommender operator would like to expand their business.

Research and production evaluation of recommender systems needs to account for how different subsets of the user population should be weighted. There is not necessarily a one-size-fits-all answer to the question of how to structure an evaluation; it is a decision that needs to be made based on the values and goals of the business or research program. Our methods and results can provide data to understand the ramifications of the decisions made about recommender evaluation.

5.2. Interaction with Popularity Bias

Popularity bias (Bellogin et al., 2011) describes the phenomenon in which offline top-N recommender evaluation gives higher scores to algorithms that favor popular items. The extent to which this is a defect in the evaluation — favoring popularity irrespective of user preference — versus an actual measurement of the effectiveness of popular recommendations is unclear; it is believed that it represents a significant deviation from ‘true’ performance, but the degree of that deviation is difficult to quantify.

From first principles, we expect popularity bias to exacerbate demographic biases: the patterns of the largest group of users will dominate the list of most-popular items, so favoring popular recommendations will also favor recommendations that are more likely to match the taste of the dominant group of users at the expense of other groups with different favorite items.

However, our empirical results do not demonstrate that effect in the data we have. Some of the demographic differences in recommender accuracy that we see, such as the ML1M gender difference, correlate with the size of the user group; others, such as LFM1K gender differences and LFM360K age differences, do not.

It is difficult to generalize about the causes of the differences we have seen from only three data sets, but it is clear that we need to look beyond popularity bias and demographic group size to understand the drivers of demographic differences in recommender performance. The consistency of the results across algorithm families, however, suggests some robustness to these effects.

Further, we have observed that applying one technique for reducing popularity bias can shift our measurements of demographic bias. This indicates tradeoffs in the measurement of different biases, so that applying the popularity bias reduction method is not a clearly correct decision.

5.3. User Retention

One of the goals of recommender systems is to engage users with the systems themselves, so that over time, users can benefit, in terms of personalization, given the availability of explicit preference data.

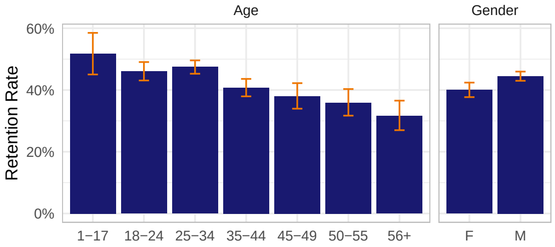


Figure 10: Retention rate for each demographic group on the ML1M data set (with 95% Wilson confidence intervals).

A way to quantify this engagement is through *retention*: do users continue using the system? The ML1M data set includes timestamps for each rating, allowing us to analyze user activity over time; we use this to measure retention and examine its relationship to demographic group. We divide user rating activity into *sessions* by considering the user to be starting a new session whenever there is a gap of at least an hour between two ratings (Halfaker et al., 2014). Figure 10 shows the retention rate (the percentage of users who returned for a second session) for each demographic group.

We observe that men have a higher retention rate than women ($p < 0.005$); in the ML1M data set, the algorithms we tested provide more accurate recommendations to men than women. While this by no means demonstrates a causal link — for one thing, we are not testing the same algorithm and implementation that MovieLens employed when these users were active — it suggests room for further exploration. The link between recommendation quality and user retention is key to the online testing employed by large-scale recommender system operators such as Netflix.

5.4. Limitations of Data

Our analysis on the ML1M data set was conducted with users' explicit feedback, the provided ratings. While this data set shows that a certain demographic group dominates its counterparts in providing ratings in the system, it does not account for implicit feedback, or the behavior of users who watched movies without necessarily providing ratings for them. To ensure that we accounted for the differences in how demographic groups prefer to provide feedback, we also performed an analysis on Last.FM based on the number of times a song was played. Our results show consistency across the different groups irrespective of the type of user feedback, i.e., implicit or explicit.

While our results highlight the need to consider disparate demographic groups when evaluating recommender systems to better account for user satisfaction, the users of MovieLens and Last.FM may not be representative of general recommender system users. Both of these systems (particularly at the time the Last.FM data was collected) appeal to experienced users who care deeply about their movies and music. Casual users are more likely to use services such as Netflix and Spotify, and may exhibit markedly different behavior and experience different recommender utility than the expert users in the data sets we examined. Unfortunately, data from more widely-used systems with sufficient attributes to look for demographic effects is difficult to find. Many widely-used data sets, such as Amazon.com and Netflix, do not contain user demographics.

5.5. Limitations of Evaluation Protocol

The fact that our results are in an entirely offline experimental setting also introduces limitations. Our data cannot distinguish whether the differences in measured performance are due to actual differences in the recommender's ability to meet users' information needs, or differences in the evaluation protocol's effectiveness at measuring that ability. While we suspect that they do reflect actual differences in recommender utility, additional study with online evaluation is needed to complement and calibrate these results, as the correlation between offline accuracy and online measures of effectiveness is often weak (Rossetti et al., 2016).

A similar concern can be raised for online protocols (Mehrotra et al., 2017), but the closer connection between online measures and long-term customer value and experience improves their external validity. However, even if our observed differences are due in significant part to limitations of the evaluation protocol, the result is still interesting: biases in the evaluation protocol for or against groups of users would impede the development of fair recommender systems. Even institutions that can carry out online evaluations use offline protocols to pre-screen algorithms prior to live deployment, and offline evaluation metrics are the basis for the objective functions in many recommender model-training processes.

5.6. Limitations of Algorithm Selection

While we have tested representatives of several key families of collaborative filtering algorithms, there are many types of algorithms that we have not considered. Two notable omissions are content-based filters, which we omitted because only one of our data sets has sufficient data to support them, and learning-to-rank recommenders, which LensKit does not yet provide.

Our evaluation methodology and open experimental scripts make it easy to re-run our analyses on additional algorithms as they become available in the underlying software.

5.7. Choice of Metric

There are many widely-used metrics that can be used to evaluate recommender systems (Gunawardana and Shani, 2009). For clarity and space, we focus our results in Section 4 on nDCG, because it considers all of a user's test items and has a good conceptual mapping to recommendation utility. We included several metrics in our experimental runs, however, and they showed similar result trends.

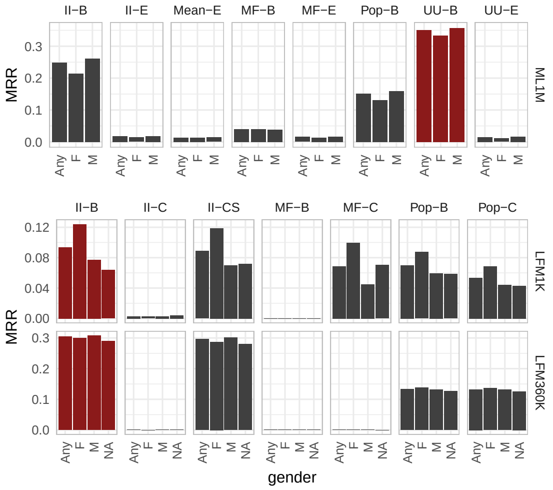


Figure 11: Algorithm performance based on Mean Reciprocal Rank for users grouped by gender.

Figures 11 and 12 show our key results from Section 4.1 with the Mean Reciprocal Rank (MRR) metric (Kantor and Voorhees, 1997). MRR measures recommendation accuracy effectiveness by taking the reciprocal of the position of the first relevant suggestion in each user's ranked list recommendations and averages this value over all users in the data set. These performance trends match those in Figures 3 and 4: (i) male users gain better utility from varied recommendation strategies than female users in ML1M and LFM360K, (ii) female users gain better utility on LFM1k data sets, and (iii) there are age differences that do not map to demographic group size. The difference in recommender system satisfaction among users of different genders is more prominent when measured by MRR than nDCG. We hypothesize that this is due to the fact that MRR penalizes recommendations that move the first relevant item (i.e., highly rated items) further down the list more heavily than nDCG, especially in long lists. On the other hand, nDCG considers the position of all relevant recommendations, along with their value to the user, instead of only observing the position of the first item. Which metric is a better measurement of usefulness depends on the precise recommendation task.

5.8. Ethical Considerations

As our work is entirely based on widely-distributed public data and we did not perform any data linking that might expose or deanonymize users in the underlying data sets, it does not place MovieLens or Last.FM users at any risk to which they have not already been ex-

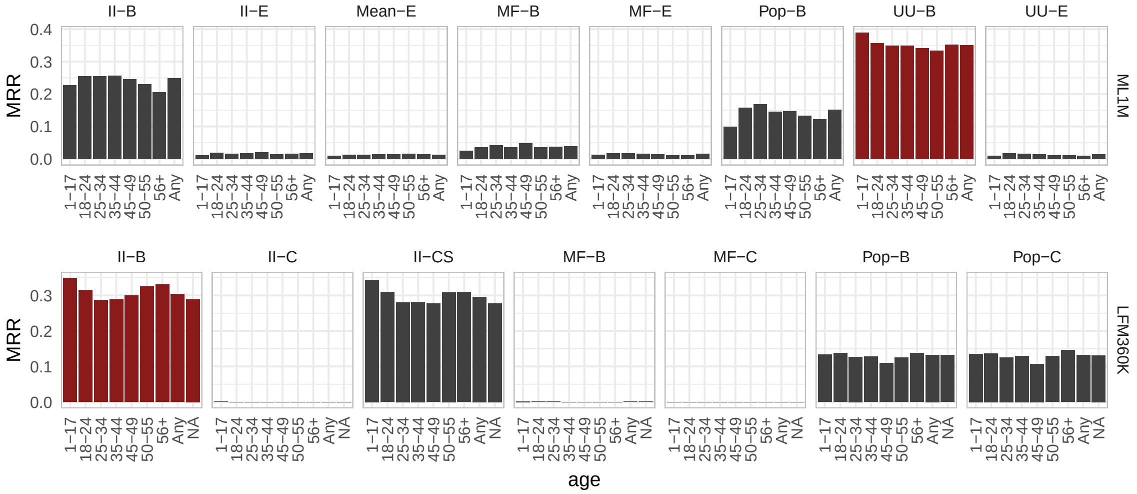


Figure 12: Algorithm performance based on Mean Reciprocal Rank for users grouped by age.

posed for years by the publication of these data sets.

6. Conclusion and Future Work

We set out to consider whether recommender systems produced equal utility for users of different demographic groups. Using publicly available data sets, we compared the utility, as measured with nDCG, for users grouped by age and gender.

Regardless of the recommender strategy considered, we found significant differences for the nDCG among demographic groups. Selecting the best algorithm from the families we tested, ML1M and LFM1K data sets showed statistically significant differences in effectiveness between gender groups while the LFM360K data set highlighted a significant effect based on user age.

The demographic effect remains when controlling for the amount of training data available for a user; it is diminished, but may not entirely disappear, when resampling the underlying data to train the recommender on a gender-balanced data set.

Notably, the effects in utility did not exclusively benefit large groups: we observed more accuracy for women on the Last.FM data, despite the lower representation of female users in the respective Last.FM data sets.

6.1. Future Work for Research

While our analysis focused on whether this effect could be found across a variety of common recommendation algorithms, the differences appear to vary from algorithm to algorithm. This suggests there is room for considering how different algorithms respond to evaluation and what characteristics contribute to more uniform utility. Analysis can also be expanded to include more families of algorithms, such as content-based recommendation and learning-to-rank techniques.

Having found this effect with age and gender, we have not yet considered intersectionality: how does recommender effectiveness vary with the interaction of multiple demographic data? Our analysis did not find that smaller groups were always disadvantaged, so more research should be done to understand *why* groups are unevenly advantaged by recommendation algorithms.

There is also room for this analysis to be repeated across other item domains. How recommender system utility compares across demographics may be especially interesting for domains like real estate, housing, and job recommendations, areas with well-documented historical discrimination.

6.2. Future Work for Industry

Given the hazards of publishing data sets which include individual user's demographic information, there are limits to the advances academia can pursue in this work. As it falls to industry to consider whether their own recommender systems provide comparable utility across demographics, so does the responsibility for publishing their results. We see the work of Mehrotra et al. (2017) as an exemplary start in this direction, although we would like to see additional details provided to ease replication of the results for other system operators.

6.3. Towards Fair Recommendation

Research on the fairness of recommender systems is just getting started, and there are many important questions to explore. We have focused on one small corner of the problem: the equity of recommender utility as experienced by different groups of users. As Burke (2017) shows, there are many more dimensions to the problem, such as the equitable treatment of content producers, as well as the distribution of non-accuracy recommendation value like diversity and serendipity.

Acknowledgments

This work was partially supported by NSF grant IIS 15-65936.

References

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

Alejandro Bellogin. *Performance prediction and evaluation in Recommender Systems: an Information Retrieval perspective.* PhD thesis, UAM, 2012.

Alejandro Bellogin, Pablo Castells, and Ivan Cantador. Precision-oriented evaluation of recommender systems: an algorithmic comparison. In *Proc. of ACM RecSys*, page 333–336. ACM, 2011. ISBN 9781450306836. doi: 10.1145/2043932.2043996. URL http://doi.acm.org/10.1145/2043932.2043996.

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

Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

Robin Burke. Multisided fairness for recommendation. *Computing Research Repository*, July 2017. URL http://arxiv.org/abs/1707.00093.

Robin Burke, Nasim Sonboli, Masoud Mansoury, and Aldo Ordonez-Gauger. Balanced neighborhoods for Fairness-Aware collaborative recommendation. In *FATREC Workshop on Fairness, Accountability and Transparency in Recommender Systems at RecSys*, 2017. URL http://scholarworks.boisestate.edu/fatrec/2017/1/3/.

O. Celma. *Music Recommendation and Discovery in the Long Tail.* Springer, 2010.

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proc. of ACM RecSys*, page 39–46. ACM, 2010. ISBN 9781605589060. doi: 10.1145/1864708.1864721. URL http://doi.acm.org/10.1145/1864708.1864721.

Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. *ACM TOIS*, 22(1):143–177, 2004.

Michael Ekstrand, John Riedl, and Joseph 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.

Michael D Ekstrand and Vaibhav Mahant. Sturgeon and the cool kids: Problems with Top-N recommender evaluation. In *Proc. of FLAIRS*. AAAI Press, 22 May 2017. URL https://md.ekstrandom.net/research/pubs/sturgeon/.

Michael D. Ekstrand and Maria Soledad Pera. The demographics of cool: Popularity and recommender performance for different groups of users. In *RecSys 2017 Poster Proceedings*, 2017.

Michael D Ekstrand, Michael Ludwig, Joseph A Konstan, and John T Riedl. Rethinking the recommender research ecosystem: reproducibility, openness, and lenskit. In *Proc. of ACM RecSys*, 2011.

Simon Funk. Netflix update: Try this at home. http://sifter.org/~simon/journal/20061211.html, December 2006. Accessed: 2010-4-8.

Mustansar Ali Ghazanfar and Adam Prugel-Bennett. A scalable, accurate hybrid recommender system. In *Proc. of IEEE WKDD*, pages 94–98, 2010.

Carlos A Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM TMIS*, 6 (4):13:1–13:19, December 2015. doi: 10.1145/2843948.

Bertram Myron Gross. *The managing of organizations: The administrative struggle*, volume 2. [New York]: Free Press of Glencoe, 1964.

Asela Gunawardana and Guy Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10:2935–2962, December 2009. ISSN 1532-4435. URL http://jmlr.org/papers/v10/gunawardana09a.html.

Aaron Halfaker, Oliver Keyes, Daniel Kluver, Jacob Thebault-Spieker, Tien Nguyen, Kenneth Shores, Anuradha Uduwage, and Morten Warncke-Wang. User session identification based on strong regularities in inter-activity time. *arXiv:1411.2878 [cs]*, November 2014. URL http://arxiv.org/abs/1411.2878.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Transactions on Interactive Intelligent Systems*, 5(4):19, 2016.

Jon Herlocker, Joseph A Konstan, and John Riedl. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Information retrieval*, 5(4):287–310, 2002.

Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM TOIS*, 22(1):5–53, 2004.

William Hill, Larry Stead, Mark Rosenstein, and George Furnas. Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 194–201, 1995. doi: 10.1145/223904.223929. URL http://dx.doi.org/10.1145/223904.223929.

Kartik Hosanagar, Daniel Fleder, Dokyun Lee, and Andreas Buja. Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation. *Management Science*, 60(4):805–823, November 2013. doi: 10.1287/mnsc.2013.1808.

Dirk Hovy. Demographic factors improve classification performance. In *ACL (1)*, pages 752–762, 2015.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM TOIS*, 20(4):422–446, October 2002. doi: 10.1145/582415.582418.

F Kamiran and T Calders. Classifying without discriminating. In *Proc. of 2nd International Conference on Computer, Control and Communication*, pages 1–6, February 2009. doi: 10.1109/IC4.2009.4909197.

Toshihiro Kamishima and Shotaro Akaho. Considerations on recommendation independence for a Find-Good-Items task. In *Proc. of Workshop on Fairness, Accountability and Transparency in Recommender Systems at RecSys*, 2017. URL http://scholarworks.boisestate.edu/fatrec/2017/1/11/.

Paul B Kantor and Ellen Voorhees. Report on the TREC-5 confusion track. In *The Fifth Text REtrieval Conference (TREC-5)*, October 1997. URL http://trec.nist.gov/pubs/trec5/t5_proceedings.html.

Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012.

Ron Kohavi, Randal M Henne, and Dan Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the HiPPO. In *Proc. of ACM KDD*, page 959–967, 2007. ISBN 9781595936097. doi: 10.1145/1281192.1281295. URL http://portal.acm.org/citation.cfm?doid=1281192.1281295.

Stefan Langer and Joeran Beel. The comparability of recommender system evaluations and characteristics of docear's users. In *Proc. of Workshop on Recommender Systems Evaluation: Dimensions and Design (REDD) at ACM RecSys*, pages 1–6, 2014.

G Linden, B Smith, and J York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003. doi: 10.1109/MIC.2003.1167344.

Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook*, pages 73–105. Springer, 2011.

Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. Auditing search engines for differential satisfaction across demographics. In *Proc. of WWW Companion*, 2017. ISBN 9781450349147. doi: 10.1145/3041021.3054197. URL https://doi.org/10.1145/3041021.3054197.

Arkadiusz Paterek. Improving regularized singular value decomposition for collaborative filtering. In *Proc. of KDD cup and workshop*, volume 2007, pages 5–8, 2007.

Michael J Pazzani. A framework for collaborative, content-based and demographic filtering. *Artificial intelligence review*, 13(5-6):393–408, 1999.

Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer, 2007.

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proc. of ACM CSCW*, pages 175–186, 1994.

Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B Kantor. *Recommender Systems Handbook*. Springer, 2015.

Marco Rossetti, Fabio Stella, and Markus Zanker. Contrasting offline and online results when evaluating recommendation algorithms. In *Proc. of ACM RecSys*, page 31–34, 2016. doi: 10.1145/2959100.2959176.

Alan Said, Till Plumbaum, Ernesto W De Luca, and Sahin Albayrak. A comparison of how demographic data affects recommendation. *UMAP*, page 7, 2011.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proc. of WWW*, pages 285–295. ACM, 2001.

Guy Shani and Asela Gunawardana. Evaluating recommendation systems. In *Recommender Systems Handbook*, pages 257–297. Springer, 2011.

Alvin Toffler. *Future shock*. Bantam, 1990.

Marshall van Alstyne and Erik Brynjolfsson. Global village or Cyber-Balkans? modeling and measuring the integration of electronic communities. *Management Science*, 51(6):851–868, June 2005. doi: 10.1287/mnsc.1050.0363.

Ingmar Weber and Carlos Castillo. The demographics of web search. In *Proc. of ACM SIGIR*, pages 523–530. ACM, 2010.

Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. *Computing Research Repository*, May 2017. URL http://arxiv.org/abs/1705.08804.