# Addressing Marketing Bias in Product Recommendations

Mengting Wan*
Airbnb, Inc.
mengting.wan@airbnb.com

Jianmo Ni
UC San Diego
jin018@eng.ucsd.edu

Rishabh Misra
Twitter, Inc.
rmisra@twitter.com

Julian McAuley
UC San Diego
jmcauley@eng.ucsd.edu

## ABSTRACT

Modern collaborative filtering algorithms seek to provide personalized product recommendations by uncovering patterns in consumer-product interactions. However, these interactions can be biased by how the product is marketed, for example due to the selection of a particular human model in a product image. These correlations may result in the underrepresentation of particular niche markets in the interaction data; for example, a female user who would potentially like motorcycle products may be less likely to interact with them if they are promoted using stereotypically 'male' images.

In this paper, we first investigate this correlation between users' interaction feedback and products' marketing images on two real-world e-commerce datasets. We further examine the response of several standard collaborative filtering algorithms to the distribution of consumer-product market segments in the input interaction data, revealing that marketing strategy can be a source of bias for modern recommender systems. In order to protect recommendation performance on underrepresented market segments, we develop a framework to address this potential marketing bias. Quantitative results demonstrate that the proposed approach significantly improves the recommendation fairness across different market segments, with a negligible loss (or better) recommendation accuracy.

## 1 INTRODUCTION

By connecting users to relevant products across the vast range available on e-commerce platforms, modern recommender systems are already ubiquitous and critical on both sides of the market, i.e., consumers and product sellers. Among recommendation algorithms used in practice, many fall under the umbrella of *collaborative filtering* [13, 17, 20, 25], which collect and generalize users' preference patterns from logged consumer-product interactions (e.g. purchases, ratings). These feedback interactions can be *biased* by multiple factors, potentially surfacing unfair (or irrelevant) recommendations

---

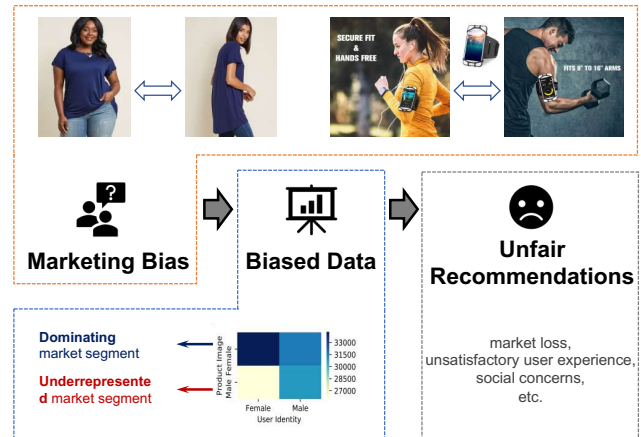*Work was done while the author was at UC San Diego.

Figure 1: Two illustrative examples on how the same product can be marketed using different human images (different body shapes, different genders). These marketing strategies could affect consumers' behavior thus resulting in a biased interaction dataset, which is commonly used as the input for modern recommender systems.

to users or items underrepresented in the input data. Such phenomena have already raised some attention from the recommender system community: a handful of types of algorithmic biases have been addressed, including selection bias [26], popularity bias [29], and several fairness-aware recommendation algorithms have been proposed [3, 6]. In this paper, we focus on a relatively underexplored factor—*marketing bias*—in consumer-product interaction data, and study how recommendation algorithms respond to its effect.

We are particularly interested in the human factors, such as the profile of the human model in a product image, reflected in a product's marketing strategies, which (as indicated in previous marketing studies) could possibly affect consumers' interactions and satisfaction [5, 11, 12]. A common hypothesis (known as 'self-congruence') is that a consumer may tend to buy a product because its public impression (in our case a *product image*), among other alternatives, is *consistent* with one's self-perceptions (*user identity*) [12]. Based on this assumption, the selection of human models for a product (as shown in Figure 1, a product can be represented by models with different body shapes or different genders) could influence a consumer's behavior. For example, a female user may be less likely to interact with an armband product which is presumably gender-neutral but marketed exclusively via 'male' images. As with many other types of bias, this could lead to underrepresentation of some niche market segments in the input data for a recommender system. Note if undesired patterns are propagated into recommendation results (e.g. even fewer male-represented products are recommended

to the potential female users), utility from both sides of the marketplace could be harmed. That is, product retailers may lose potential consumers while users may be struggling to find relevant products. As a consequence, serious ethical and social concerns could be raised as well.

In this work, we seek to understand 1) if such a marketing bias exists in real-world e-commerce datasets; 2) how common collaborative filtering algorithms interact with these potentially biased datasets; and 3) how to alleviate such algorithmic bias (if any) and improve the market fairness of recommendations. We conclude our contributions as follows.

- We collect and process two e-commerce datasets from ModCloth and Amazon. Then we conduct an observational study to investigate the relationship between interaction feedback and product images (reflected in the selection of a human model) as well as user identities. Different types of correlations in varying degrees can be observed in these two datasets.

- We implement several common collaborative filtering algorithms and study their responses to the above patterns in the input data. For most algorithms, we find 1) systematic deviations across different consumer-product segments in terms of rating prediction error, and 2) notable deviations of the resulting recommendation outputs from the real interaction data.

- Note that as the marketing bias could be intricately entangled with users' intrinsic preferences, our goal in this work is not to pursue the absolute parity of recommendations (e.g. keep recommending products represented by human images which were constantly unfavored by a user). Rather, we expect a fair algorithm is supposed *not to worsen the market imbalance* in interactions. We thus propose a fairness-aware framework to address it by calibrating the parity of prediction *errors* across different market segments. Quantitative results indicate that our framework significantly improves recommendation fairness and provides better accuracy-fairness trade-off against several baselines.

## 2 RELATED WORK

This work is partially motivated by the well-known 'self-congruity' theory in marketing research, which is defined as the match between the product/brand image and the consumer's true identity and the perception about oneself [5, 11, 12]. Many previous marketing studies focus on assessing this theory by quantifying and validating it through statistical analysis on a small amount purchase transaction data or the feedback in questionnaires [18, 21, 27, 28]. Following self-congruity theory, products can be advertised in a way to match their target consumers' images thus establishing product stereotypes [10]. Our work is distinguished with these studies from a more computational perspective, by identifying and studying the potential marketing bias for recommender systems on large-scale e-commerce interaction datasets.

Our analysis is related to previous work which examines particular types of biases in real-world interactions and their effects in recommendation algorithms, including the popularity effect and catalog coverage [14], the bias regarding the book author gender for book recommenders [7], and the herding effect in product ratings [31].

| | ModCloth | Electronics |
|---|---|---|
| #review | 99,893 | 1,292,954 |
| #item | 1,020 | 9,560 |
| #user | 44,783 | 1,157,633 |
| time span | 2010-2019 | 1999-2018 |
| bias type | body shape | gender |
| product image | Small (838) Small&Large (182) | Female (4,090) Female&Male (2,466) Male (3,004) |
| user identity | Large (9,395) Small (30,140) N/A (5,248) | Female (71,043) Male (61,350) N/A (1,025,240) |

**Table 1: Basic statistics of the *ModCloth* and *Electronics* datasets.**

Another closely related line of work includes developing evaluation metrics and algorithms to address fairness issues in recommendations. 'Unbiased' recommender systems with missing-not-at-random training data are developed by considering the propensity of each item [15, 26]. A fairness-aware tensor-based algorithm is proposed to address the absolute statistical parity (i.e., items are expected to be presented at the same rate across groups) [32]. Several fairness metrics and their corresponding algorithms are proposed for both pointwise prediction frameworks [6, 30] and pairwise ranking frameworks [3]. Methodologically, these algorithms can be summarized as reweighting schemes where underrepresented samples are upweighted [6, 15, 26] or schemes where additional fairness terms are added to regularize the model [1, 3, 30].

Note that most of the above studies focus on bias and fairness on one side of the market only (i.e., either user or producer). Our concern about marketing bias is that it could affect fairness for both consumers and product providers. Without global market fairness in mind, the imbalance of the consumer-product segment distribution could be exacerbated through the deployment of recommendation algorithms. Multi-sided fairness is addressed by Burke *et al* [3] by considering C(onsumer)-fairness and P(rovider)-fairness. Trade-off between accuracy and fairness in two-sided marketplaces is further explored and a counterfactual framework is proposed to evaluate different recommendation policies without extensive A/B tests [22]. However the CP-fairness condition where fairness is protected for both sides *at the same time* still remains an open question.

## 3 DATA COLLECTION AND PREPROCESSING

We introduce two real-world e-commerce datasets collected from a women's clothing website *ModCloth*[1] and the *Electronics* category on *Amazon*.[2] These datasets enable us to study the marketing bias induced by the selection of a human model with respect to *body shape* for clothing products, and investigate the effects from the *gender* of human models for electronics products. Detailed information about these datasets can be found in Table 1. Datasets in this paper are available at https://github.com/MengtingWan/marketBias.

Note that our datasets are not perfect, e.g. errors and selection bias can be introduced via scraping, parsing and processing, control

---

[1]https://www.modcloth.com/
[2]https://www.amazon.com/

of several confounding factors including inventory status, (etc.). Our intention here is neither to make any normative claims regarding the distributions in the above two applications, nor draw any causal conclusions. Rather, we simply describe the current state of these datasets and *study how recommendation algorithms interact with these data.*

## 3.1 ModCloth

***ModCloth*** is an e-commerce website which sells women's clothing and accessories. One unique property of this data is that many products include two human models with different body shapes (as shown in Figure 1) and measurements of these models. In addition, users can optionally provide the product sizes they purchased and fit feedback ('Just Right', 'Slightly Larger', 'Larger', 'Slightly Smaller' or 'Smaller') along with their reviews. Therefore we focus on the dimension of *human body shape* as the source of marketing bias in this dataset.

**Product Image Group (Body Shape).** We start with the clothing products included in an existing public dataset [23], re-scrape their landing pages, collect related model size measurements and all review ratings. We normalize their product sizes as 'XS', 'S', 'M', 'L', 'XL', '1X', '2X', '3X' and '4X' according to the provided size charts.[3] Products with only one human model wearing a relatively small size ('XS', 'S', 'M' or 'L') are labeled as the 'Small' group while products with two models (an additional model wearing a plus-size: 1X', '2X', '3X' or '4X') are referred as the 'Small&Large' group.

**User Identity Group (Body Shape).** We then calculate the average size each user purchased and classify users into 'Small' and 'Large' groups based on the same standard as the product body shape image.

We observe that all products offer the complete spectrum of sizes, while 70% of these products are interacted with by at least one user from the 'Large' group and 97% are interacted with by the 'Small' group. Thus we conclude that most users are able to consume most products at some point within the time frame of our dataset.

Ultimately we collect nearly 100K reviews about 1,020 clothing products from 44,783 users, where around 90% of users can be matched to the above identity groups.

## 3.2 Electronics

***Electronics*** is another review dataset collected from the *Electronics* category on Amazon with *Clothing* as an auxiliary category. This dataset is built on top of the public *Amazon 2018 Dataset* [24] and further processed to facilitate the research goals in this paper. We regard the *gender* as the target marketing bias on this dataset.

**Product Image Group (Gender).** In the *Amazon 2018 Dataset*, we keep all pictures associated with electronic products[4] and run human model detection through an industrial body/face detection API provided by Face++.[5] The results include whether any human bodies/faces are included in the pictures, as well as gender predictions of these detected models. We only keep products where human models are detected in their associated pictures and treat them as

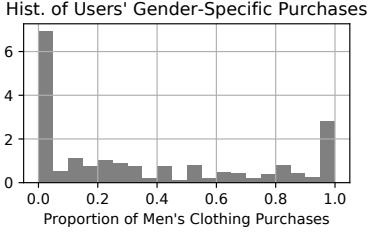| Term/Symbol | Description |
|---|---|
| product image | the public impression of a product; attributes of the human models included in the product pictures are used in this work, e.g. body shape, gender |
| user identity | the perception of oneself; we use the same dimension of attribute as in product image |
| $m, n$ | user identity group, product image group, e.g. female/male |
| $M, N$ | the number of possible user identity groups and product image groups |
| $s_{u,i}, r_{u,i}, e_{u,i}$ | predicted user $u$'s preference score on product $i$, user $u$'s rating score on product $i$, prediction error $e_{u,i} = s_{u,i} - r_{u,i}$ |
| $\mathcal{U}_m, \mathcal{I}_n$ | the user set with the same identity $m$, the item set with the same product image $n$ |
| market segment $(m, n)$ | the market defined for users with the same identity $m$ on products with the same type of image $n$ |
| $\mathcal{D} = \{r_{u,i}\|u,i\}$ | the complete interaction data |
| $\mathcal{D}_{m,n}, \mathcal{D}_m, \mathcal{D}_n$ | interactions within the market segment $(m,n)$, $\mathcal{D}_m = \bigcup_n \mathcal{D}_{m,n}, \mathcal{D}_n = \bigcup_m \mathcal{D}_{m,n}$ |

**Table 2: Important terms and notation.**

three types of product gender image based on the selection of these human models: 'Female' (only female models are included), 'Male' (only male models are included) and 'Female & Male' (both female and male models are detected, not necessarily in the same picture).

We then involve 3 human labelers to conduct validations on this dataset, where label conflicts are resolved by majority voting. 3,000 randomly sampled pictures are manually labeled regarding 1) if they notably include human models; 2) the gender image from 'Female Exclusive', 'Male Exclusive' or 'Both Female & Male' (if multiple models are included in a single picture). We evaluate the human model detection results from the API based on these labels and find a high precision (96%) regarding the human model detection but a relatively low recall (53%). Note in our setting we are happy to discard ambiguous cases (sacrifice some recall) for the sake of high precision. We later randomly sample 100 products and manually decide if these products preserve any gender constraints based on their descriptions. Although 4 out of 100 products exhibit gender implications,[6] we don't find any strict constraints which prevent the unfavorable user identity group from consuming these products.

**User Identity Group (Gender).** Unfortunately, gender identities of Amazon users are not directly accessible. We thus leverage users' interactions with *Clothing* products in the *Amazon 2018 Dataset* to access their gender identities, where most products are explicitly classified into *Women's Clothing* or *Men's Clothing*. As shown in the figure below, we find a clear bimodal distribution of purchase frequency towards gender-specific clothing products.

---

[3]e.g. https://www.modcloth.com/size-guide.html
[4]All products attached to the 'Men' or 'Women' categories are removed.
[5]https://www.faceplusplus.com/

[6]e.g. https://www.amazon.com/gp/product/B00HX19EDI

Hist. of Users' Gender-Specific Purchases

| | ModCloth | | | Electronics | | |
|---|---|---|---|---|---|---|
| | $\chi^2$ | p-value | #reviews | $\chi^2$ | p-value | #reviews |
| all | 158.7 | <0.001 | 91,526 | 581.8 | <0.001 | 174,124 |
| <=2014 | 0.5 | 0.466 | 25,383 | 151.0 | <0.001 | 49,699 |
| 2015 | 66.7 | <0.001 | 20,241 | 172.7 | <0.001 | 46,891 |
| 2016 | 70.8 | <0.001 | 21,239 | 96.4 | <0.001 | 43,907 |
| >=2017 | 29.0 | <0.001 | 24,663 | 120.8 | <0.001 | 33,627 |

**Table 3: Results from $\chi^2$ test of the two-way contingency tables on *ModCloth* and *Electronics*.**

| | User Identity | | |
|---|---|---|---|
| **Product Image** | Small | Large | All |
| Small | 31,800 (+754.98) | 7,038 (-754.98) | 38,838 |
| Small&Large | 41,361 (-754.98) | 11,327 (+754.98) | 52,688 |
| All | 73,161 | 18,365 | 91,526 |

**(a) ModCloth**

| | User Identity | | |
|---|---|---|---|
| **Product Image** | Female | Male | All |
| Female | 34,259 (+1,472.89) | 31,587 (-1,472.89) | 65,846 |
| Female&Male | 26,478 (+880.88) | 24,930 (-880.88) | 51,408 |
| Male | 25,963 (-2,353.77) | 30,907 (+2,353.77) | 56,870 |
| All | 86,700 | 87,424 | 174,124 |

**(b) Electronics**

**Table 4: Contingency tables of the frequency distribution of product images and user identities on *ModCloth* and *Electronics*. Deviations ($f_{m,n} - \mathbb{E}f_{m,n}$) from the expected frequency values are provided in parentheses.**

We discard 'ambiguous' users whose men's (or women's) clothing purchase frequencies fall into 40%-60%, and identify the remaining users as 'Female' (69%) or 'Male' (31%). Finally 11% of total users in the *Electronics* category can be matched to these identities and 53% of them are identified as 'Female'.

After removing products without any human models, we are still able to obtain a large-scale dataset containing around 1.3M rating scores across 9,560 electronics products from 1.1M users. Note that although the inferred product gender image and user identity are not as precise as in *ModCloth*, this dataset is dramatically different from *ModCloth* regarding its scale and sparsity. In contrast to the relationship between a user's interactions with clothing products and a dimension of *human body shape*, we speculate that *gender* is intuitively less relevant to the intrinsic qualities of most products in *Electronics*; thus its effects on users' interactions are possibly more likely to come from marketing bias.

## 4 STATISTICAL ANALYSIS

We split a consumer's product preference into two dimensions: 1) the user's preference in terms of willingness to consume (purchase) a product; and 2) the user's satisfaction feedback (e.g. ratings) on the consuming experience. We then conduct observational studies on the *ModCloth* and *Electronics* datasets to address marketing bias across the above two dimensions.

- We first investigate if there is a bias introduced by a particular marketing effect in a consumer's *product selection* process. Specifically, we examine if a correlation exists between product image and user identity in terms of interaction frequency in our datasets.
- Then we study *consumer satisfaction* regarding the purchased products as a function of product image, user identity, and their second-order interactions. These consumer feedback signals include rating scores on *ModCloth* and *Electronics*, as well as the binarized fit feedback (i.e., if the clothing product fits the user) on *ModCloth*.

Important terms and notation throughout the paper are included in Table 2.

### 4.1 Product Selection vs. Marketing Bias

Because of the constraint of conducting real-world experiments with random assignments, we instead address marketing bias in product selection by analyzing the association between product image and user identity in observed data with respect to interaction frequency. Our null hypothesis is that product image and user identity are statistically independent. Given this assumption, we expect to see lower deviations of their observed frequencies and the

marginally expected values. Therefore the following Pearson's Chi-Squared Test Statistic can be used to test the association between these two variables in terms of frequency [8]:

$$\chi^2 = \sum_{m,n} \frac{(f_{m,n} - \mathbb{E}f_{m,n})^2}{\mathbb{E}f_{m,n}}, \qquad (1)$$

where $m$ and $n$ represent a user identity group and a product image group respectively, $f_{m,n}$ is the observed number of interactions in the market segment $(m, n)$ and $\mathbb{E}f_{m,n} = \frac{(\sum_{m'} f_{m',n})(\sum_{n'} f_{m,n'})}{(\sum_{m',n'} f_{m',n'})}$ represents its expectation. The null hypothesis will be rejected (i.e., the association between two variables exists in terms of frequency) if an extremely large $\chi^2$ is obtained (i.e., small $p$-value).

To further separate the potential marketing bias from trending effects, we conduct association tests on the complete interaction data as well as interactions within different time spans. Test results are included in Table 3, where we find all $p$-values are smaller than 0.001 except for the test on interaction data before 2014 on *ModCloth*. These results may imply the existence of the association between product image and user identity in consumers' product selections.

In Table 4, we provide contingency tables of the frequency distribution of different market segments and their deviations from

| | ModCloth | | | | Electronics | |
| --- | --- | --- | --- | --- | --- | --- |
| | Rating | | Fit | | Rating | |
| | F-stat | p-value | F-stat | p-value | F-stat | p-value |
| product | 171.9 | <0.001 | 293.1 | <0.001 | 62.6 | <0.001 |
| user | 46.3 | <0.001 | 402.4 | <0.001 | 3.5 | 0.061 |
| user×product | 30.7 | <0.001 | 0.0 | 0.997 | 0.9 | 0.404 |

**Table 5: Results from two-way analysis of variance (ANOVA) on *ModCloth* and *Electronics*.**

expected values ($f_{m,n} - \mathbb{E}f_{m,n}$). We observe generally more interactions than expected on the consumer-product segments where users' identities match the product images ('self-congruity'), while several market segments are underrepresented in the data. For example, ('Large' user, 'Small' product) on ***ModCloth*** and ('Female' user, 'Male' product) on ***Electronics*** have smaller market sizes compared with other market segments.

## 4.2 Consumer Satisfaction vs. Marketing Bias

Next we investigate consumer satisfaction as a function of product image and user identity through a standard statistical technique: two-way analysis of variance (**ANOVA**) [16]. We use rating scores to represent users' satisfaction regarding the overall quality of their consuming experience on both ***ModCloth*** and ***Electronics***. For ***ModCloth***, we also study consumer satisfactions with respect to their fit feedback (where 'Just Right' is regarded as positive while all others are regarded as negative). The two-way **ANOVA** model can be formulated as

$$consumer\ satisfaction \sim product + user + product \times user,$$

where the null hypotheses of our tests include

(a) the average consumer satisfaction is equal across different product image groups;

(b) the average consumer satisfaction is equal across different consumer identity groups;

(c) there is no interaction effect between product groups and consumer groups with respect to satisfaction.

Given these assumptions, we may expect a lower variance of average satisfactions across different groups (*between-group variation*) compared with the summation of satisfaction variations within each group (*within-group variation*). Therefore, the standard **F-statistic**, defined as the between-group variation divided by the within-group variation [16], can be applied to evaluate the correlations.

Results from statistical tests are included in Table 5. The heatmaps of sample means within market segments and their 95% confidence intervals are provided in Figure 2. We observe that users' rating scores are significantly different across market segments on ***ModCloth***. For example 'Large' users provide lower ratings on 'Small' products (Figure 2a). Although users' fit feedback differs across product groups and user groups (hypothesis (a) and (b) are rejected in Table 5), their association regarding fit feedback is negligible (results for 'user×product' in Table 5). According to Figure 2b, we find clothing products in the ***ModCloth*** dataset generally fit better on 'Small' users, and those products represented by human models with different body shapes ('Small&Large') tend to obtain better fit feedback. Although the 'self-congruity' pattern is significant
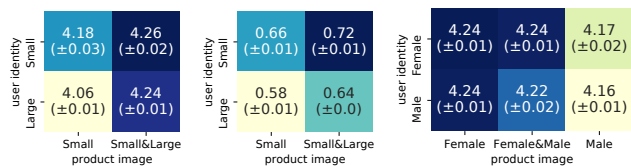


(a) ModCloth (Rt.)     (b) ModCloth (Fit)     (c) Electronics (Rating)

**Figure 2: Heatmaps of sample means within market segments regarding (a) rating scores on *ModCloth*, (b) fit feedback on *ModCloth* and (c) rating scores on *Electronics*.**

in the product selection process on ***Electronics*** (see Table 3, Table 4b), the interaction between product 'gender' and user gender is insignificant with respect to users' rating scores (user×product in Table 5).

## 4.3 Summary of Observations

We summarize insights obtained from the above statistical analysis as follows:

- The association between product image and user identity is consistently significant in terms of frequency distribution, implying the existence of marketing bias in the collected interaction datasets. The 'self-congruity' pattern is also observable, i.e., consumers may generally tend to interact with products with similar impressions as their identities. Such an association notably causes underrepresentation of certain market segments.

- The relationship between consumer satisfaction and marketing factors is rather complicated. We observe rating disparities across product groups and user groups, while the existence of their interaction effect depends on the type of product and the type of satisfaction measure. We find a similar 'self-congruity' pattern for rating scores on ***ModCloth*** while the 'user×product' term remains insignificant in the other two testing scenarios.

## 5 MARKET-FAIRNESS OF RECOMMENDER SYSTEMS

From the above analysis, we have confirmed that our interaction data is correlated to (and possibly affected by) marketing strategies used by product retailers (i.e., selections of human models). Our next step is to study if (and how) this marketing bias is propagated by algorithms from input data to recommendation results.

**Problem Setting.** In this study, we focus on recommendation algorithms trained on explicit feedback (i.e., rating scores). The primary predictive task is formulated as a *rating prediction problem*: rating scores ($r_{u,i}$) are assumed to reflect users' preferences over products, and algorithms are trained to generate users' product preference scores ($s_{u,i}$) which approximate these ratings.

Unlike previous studies [3, 4, 30, 32] which focus on evaluating and protecting the fairness of a single side (user or product) of recommender systems, in the context of marketing bias, we are particularly interested in the global market fairness of the recommendations, i.e., user-fairness and product-fairness need to be protected at the same time. Specifically we describe the market fairness in the explicit feedback setting along two dimensions.

(a) Averaged errors of rating predictions from a recommendation algorithm across different consumer-product market segments are expected to be equal.

(b) The distribution of market segments in terms of frequency within recommended interactions are expected to be consistent with the distribution within the real interaction data.

**Rating Prediction Fairness.** We notice that the first market fairness description is indeed consistent with the null hypothesis of a one-way **ANOVA** test about the association between prediction errors ($e_{u,i} = s_{u,i} - r_{u,i}$) and market segments ($(m,n)$). That is, with the assumption that average prediction errors from a fair algorithm are supposed to be irrelevant to market segments, we expect to observe a lower variation of average errors across market segments (*bewteen-segment variation*) compared to the error variations within each segment (*within-segment variation*). Specifically these variations can be defined as

between-segment var.: $\quad V^{(market)} = \dfrac{1}{|\mathcal{D}|} \sum_{m,n} |\mathcal{D}_{m,n}| \left( \bar{e}_{m,n,\cdot} - \bar{e} \right)^2$

within-segment var.: $\quad U^{(market)} = \dfrac{1}{|\mathcal{D}|} \sum_{m,n} \sum_{\substack{u \in \mathcal{U}_m, \\ i \in \mathcal{I}_n}} \left( e_{u,i} - \bar{e}_{m,n,\cdot} \right)^2$

where $\bar{e}_{m,n,\cdot}$ denotes the sample mean of prediction errors within the market segment $(m,n)$; $|\mathcal{D}_{m,n}|$ represents the number of interactions included in a consumer-product segment $(m,n)$; $|\mathcal{D}|$ denotes the total sample size.

To ensure a tractable distribution for significance testing, the above two terms are corrected by their degrees-of-freedom and the following **F-statistic** can thus be calculated:

$$\mathcal{F}^{(market)} = \frac{V^{(market)} \ / \ (M \times N - 1)}{U^{(market)} \ / \ \underbrace{(|\mathcal{D}| - M \times N)}_{\text{correction of degree of freedom}}}. \qquad (2)$$

Then we obtain a fairness evaluation metric to evaluate a global parity of prediction errors across different consumer-product market segments, where lower $\mathcal{F}$ indicates better rating prediction fairness.

**Product Ranking Fairness.** We further investigate the fairness of the product ranking performance from recommendation algorithms. For each user, we rank all products based on the predicted preference scores $s_{u,i}$ and regard the top-ranked $K$ items as recommended products. By gathering users and the recommended products, we are able to obtain the frequency distribution of market segments within these predicted interactions $p_{m,n} \sim \mathcal{P}$. We regard the frequency distribution of market segments in the real interactions $q_{m,n} \sim \mathcal{Q}$ as the reference distribution, and evaluate the deviation of $\mathcal{P}$ from $\mathcal{Q}$ using the following **KL-divergence** [19]:

$$D_{KL}(\mathcal{P}|\mathcal{Q}) = \sum_{m,n} p_{m,n} \log \left( \frac{p_{m,n}}{q_{m,n}} \right). \qquad (3)$$

We use this metric to evaluate the product ranking fairness. Lower $D_{KL}$ indicates better fairness.

## 6 A FAIRNESS-AWARE FRAMEWORK

A common optimization criterion for model-based collaborative filtering algorithms in the explicit feedback setting is based on **MSE**, i.e., minimizing the following loss function

$$\mathcal{L} = \sum (s_{u,i} - r_{u,i})^2. \qquad (4)$$

A popular choice to model the preference score $s_{u,i}$ is through matrix factorization [17]

$$s_{u,i} = b_0 + b_i + b_u + \langle \boldsymbol{\gamma}_i, \boldsymbol{\gamma}_u \rangle. \qquad (5)$$

In Eq. (5), $b_0$ is the global intercept, $b_i$ and $b_u$ are item-specific and user-specific offsets, $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_u$ are $d$-dimensional embeddings to capture items' latent properties and users' latent preferences on these dimensions.

**Error Correlation Loss.** Following previous work using the regularizing schemes [1, 3, 30], we propose a fairness-aware framework by considering an error correlation loss to regularize systematic error biases on the market:

$$\mathcal{L}^* = \sum (s_{u,i} - r_{u,i})^2 + \alpha \mathcal{L}_{corr.}, \qquad (6)$$

where $\mathcal{L}_{corr.}$ is an additional term to regularize the correlation between prediction errors $e_{u,i}$ and the distribution of market segments $(m,n)$. $\alpha$ is a hyperparameter to control the trade-off between prediction accuracy and this correlation penalty term.

In practice, we consider the following form by relaxing the evaluation metric Eq. (2):

$$\mathcal{L}_{corr.} = \kappa^{(u.)} \overbrace{\frac{V^{(u.)}}{U^{(u.)}}}^{\text{error parity on user identity}} + \kappa^{(p.)} \underbrace{\frac{V^{(p.)}}{U^{(p.)}}}_{\text{error parity on product image}} + \kappa^{(market)} \overbrace{\frac{V^{(market)}}{U^{(market)}}}^{\text{error parity on market segments}}, \qquad (7)$$

where $V^{(u.)}, U^{(u.)}, V^{(p.)}, U^{(p.)}$ can be implemented by merging market segments within the same type of user identity groups or product image groups. Note that the three error parity terms in Eq. (7) can be regarded as simplified implementations of the fairness metric in Eq. (2). $\kappa^{(u.)}, \kappa^{(p.)}, \kappa^{(market)} \in \{0,1\}^3$ are binary hyperparameters to instantiate different forms of correlation loss. For example, a selection of $(\kappa^{(u.)}, \kappa^{(p.)}, \kappa^{(market)}) = (1, 0, 0)$ represents that we only penalize the correlation between prediction errors and user identity groups.

## 7 EXPERIMENTS

We conduct experiments on the collected *ModCloth* and *Amazon* datasets to evaluate the recommendation performance and the market fairness as described in Section 5.

**Baselines.** The following standard algorithms are considered:
- **itemCF**, an item-based collaborative filtering algorithm [20, 25];
- **userCF**, a user-centric collaborative filtering method [13];
- **MF**, the matrix factorization method [17], where the value of the preference prediction $s_{u,i}$ is unbounded;
- **PoissonMF**, a hierarchical Bayesian framework where the preference factorization is linked to the rating score through a Poisson distribution, so that the preference score $s_{u,i}$ is bounded as a positive value [9].

By studying the recommendation outputs from these methods, we evaluate how standard collaborative filtering algorithms respond to the marketing bias in the input data.

We implement our proposed framework (**MF (corr.error)**), where $s_{u,i}$ is factorized using matrix factorization. By comparing its performance with the above methods (especially **MF**), we evaluate if the rating prediction and the product ranking fairness can be improved without losing much accuracy by adding the proposed correlation loss. Besides, we consider another two fairness-aware alternatives:

- **MF (corr.value)**, a method similar to **MF (corr.error)** except that $\mathcal{L}_{corr.}$ is implemented as the correlation between the predicted rating *values* $s_{u,i}$ and the market segments. By comparing **MF (corr.error)** with it, we evaluate the effectiveness of controlling the parity of prediction errors instead of the absolute statistical parity of prediction values.

- **MF (reweighted)**, a method where the loss function is reweighted based on the sizes of market segments in the training data. We also consider the following generic form of the loss function:

$$\mathcal{L} = \frac{\kappa^{(u.)}}{M} \sum_m MSE_m + \frac{\kappa^{(p.)}}{N} \sum_n MSE_n + \frac{\kappa^{(market)}}{MN} \sum_{m,n} MSE_{m,n}.$$

By comparing it with other baselines, we study if the marketing bias can be alleviated by simply increasing the weights of underrepresented segments in the training data.

For all above methods, we primarily evaluate their rating prediction accuracy through **MSE** and **MAE**, and rating prediction fairness in terms of the **F-statistic** (Eq. (2)). We also evaluate their recommendation accuracy through **AUC** and **NDCG**, and the product ranking fairness in terms of **KL-divergence** (Eq. (3)).

**Experimental Details.** We use the following rules to split interactions into train/validation/test sets: for users with at least two reviews, their most recent ratings are regarded as a test set; for users with at least three reviews, their second-to-last ratings are used for validation; the remaining interactions are used for training. We apply the same analysis on both training and test sets as in Section 4, and find similar patterns except that fewer female users (40%) are included in the test set of *Electronics*.[7]

We use the **ADAM** optimizer [2] with a learning rate of 0.001, a batch size of 512 and a fixed dimensionality of the latent embeddings in all model-based methods ($d = 10$). An $\ell_2$ regularizor is applied on all model-based methods, where $\lambda$ is selected from $\{0.01, 0.1, 1.0, 10\}$. The accuracy-fairness trade-off $\alpha$ is chosen from $\{0.5, 1.0, 5.0, 10.0\}$. All hyperparameters are selected based on the recommendation accuracy[8] on the validation set. For fairness-aware methods, we search hyperparameters $\kappa = (\kappa^{(u.)}, \kappa^{(p.)}, \kappa^{(market)})$ from $\{(1,0,0),(0,1,0),(0,0,1),(1,1,0),(1,1,1)\}$. For each $\kappa$, we first decide all other hyperparameters based on their recommendation accuracy, then select $\kappa$ which yields the fairest recommendation results on the validation set. For each user, the top-10 ranked products are regarded as recommended items. Reviews in the test set where rating scores are larger than 3 are considered as reference

---

[7] 85% female users (vs. 73% male users) have only one review in our entire dataset.
[8] **MSE** for rating prediction accurcy and **NDCG** for product ranking.



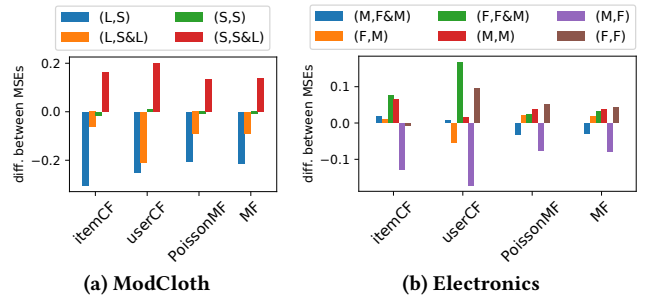**(a) ModCloth**    **(b) Electronics**

**Figure 3: Differences between the out-segment MSEs and the in-segment MSEs. Market segments are sorted based on their market sizes in the training data.**

interactions for the ranking task. All results are reported on the test set.

## 7.1 How does a standard collaborative filtering algorithm respond to biased input data?

We report the above mentioned rating prediction and product ranking metrics on *ModCloth* and *Electronics*, regarding both accuracy and fairness, in Table 6. We first investigate standard recommendation methods without any explicit fairness controls (i.e., **itemCF**, **userCF**, **PoissonMF** and **MF**). We observe that most methods yield biased prediction results on both datasets according to the **F-statistic**-based significance test. Although we find seemingly fair prediction errors from **userCF**, it actually produces a much larger **MSE** (as well as worse product ranking results) compared to other methods.

We further calculate the differences between the out-segment **MSE**s and the in-segment **MSE**s for these algorithms. Given a market segment $(m, n)$, we have

$$diff_{m,n} = MSE_{u \notin \mathcal{U}_m \text{ or } i \notin \mathcal{I}_n} - MSE_{u \in \mathcal{U}_m \text{ and } i \in \mathcal{I}_n}. \quad (8)$$

$diff_{m,n} > 0$ indicates, for an algorithm, the market segment $(m, n)$ is more predictable (smaller **MSE**) than the interactions outside it. These differences are displayed in Figure 3, where the market segments are sorted based on their training sizes. We observe an overall trend that all algorithms generally tend to favor the dominating market segments (e.g. 'Small' users on 'Small&Large' products in *ModCloth*) in varying degrees. We find the correlation between the predictibility and the market segment size is more prominent on *ModCloth* but rather complicated on *Electronics*. However, by cross matching Figure 3 and the contingency table Table 4b, we find that the trend correlates to the deviations of the real market size and the expected market size: the consumer-product segments ('Female', 'Male'), ('Male', 'Female') and ('Male', 'Female&Male') are underrepresented based on this difference ($f_{m,n} - \mathbb{E}f_{m,n} < 0$), also generally unfavored by the recommendation algorithms.

We display the distributions of market segments within positive interactions where rating scores are larger than 3 and the recommended top-10 products from these algorithms in Figure 4. Compared with the distributions in real interactions (the 'data' columns in Figure 4), we can observe the deviations of recommendation results from most algorithms, particularly **itemCF** and **userCF** on
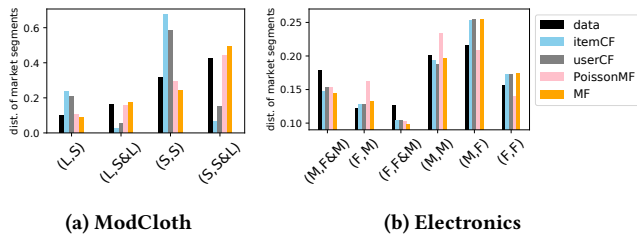
**(a) ModCloth**          **(b) Electronics**

**Figure 4: Distribution of market segments within test data (positive interactions only) and within recommendations. Market segments are sorted based on their sizes in training data.**

*ModCloth*. However, systematic patterns about how these deviations correlate to the sizes of different market segments in the training data are not observed.

## 7.2 Can recommendation fairness be improved by applying the correlation loss?

In Table 6, we further compare the results from fairness-aware algorithms in group (b) to the standard algorithms in group (a), particularly **MF**. To better visualize the trade-off between recommendation accuracy and market fairness, we present scatter plots of an accuracy metric and a fairness metric on both datasets in Figure 5. We notice the proposed method with error correlation loss **MF (corr.error)** generally provides better rating and ranking fairness (lower **F-statistic** and **KL-divergence**) than standard **MF**, without trading-off much recommendation accuracy. An interesting finding is the combination selection $\kappa$ on the validation set is consistent with our analysis in Table 5: the complete correlation loss ($\kappa = (1, 1, 1)$) is selected for *ModCloth* and the addition of product and user correlation ($\kappa = (1, 1, 0)$) is selected for *Electronics*.

We find the reweighting scheme also benefits the fairness metrics, particularly in the product ranking setting. One surprising finding is that by applying the error correlation loss, a significant performance gain in terms of product ranking accuracy (**AUC** and **NDCG**) can be obtained on *Electronics*. A possible reason could be that *Electronics* is an extremely sparse dataset where algorithms like **MF** may struggle to converge to an ideal local optimum. The fairness-aware correlation loss, however, could help regularize the training process.

## 8 CONCLUSIONS AND FUTURE WORK

We conclude our work and summarize our findings as follows:

- We investigated a potential source of bias—marketing bias—in the form of the association between interaction feedback, product image and user identity, on two real-world e-commerce datasets. Through observational studies, the inter-correlations between these factors can be confirmed and the 'self-congruity' patterns are noticeable in the product selection process, which eventually results in the underrepresentation of some market segments.

- We focused on market fairness and investigated how standard collaborative filtering algorithms react to this biased input data. We found such a bias can be propagated to the recommendation outcomes in varying degrees.



**(a) ModCloth (Rating)**          **(b) Electronics (Rating)**



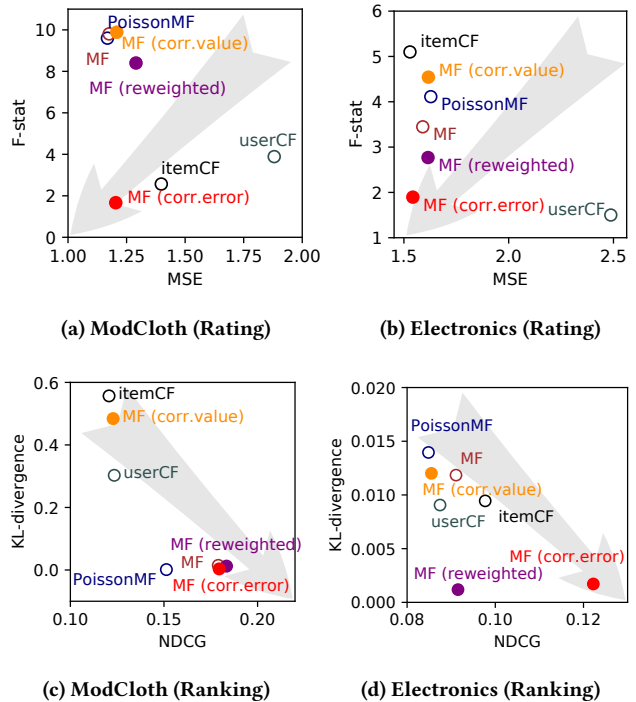**(c) ModCloth (Ranking)**          **(d) Electronics (Ranking)**

**Figure 5: Scatter plots for accuracy-fairness trade-off from different algorithms. Shaded arrows indicate the most ideal direction: higher accuracy, better fairness.**

- We developed an error correlation framework, which explicitly calibrates the equity of prediction errors across different market segments. Experimental results demonstrate that by applying this correlation loss, a superior accuracy-fairness trade-off can be achieved.

This work is a first step to approach the potential marketing bias in machine learning systems. We also wish to address several limitations of our data and methods, and to provide potential research directions.

- **Data.** We study marketing bias by formulating it as the relationship between the human model images of products and user identities. Multiple marketing factors (e.g. product descriptions, social media advertisement contents) can also be considered. Binary gender identities are inferred in our *Electronics* dataset, which is limited to represent user identities that are not exclusively masculine or feminine, e.g. users who don't always purchase products corresponding to their own identities, or those who identify themselves outside the binary definition.

- **Analysis.** We collect *ModCloth* and *Electronics* as logged interactions where many confounders (inventory status, the observability of each product, potential biases introduced in the scraping and preprocessing stage, etc.) exist and are difficult to be disentangled. Although the inter-correlation between product image and user identity is observed in these datasets, we cannot draw any causal conclusions without controlling some notable confounding factors. Therefore another direction to validate (or more fundamentally address) this marketing bias is to conduct

| | Method | ModCloth | | | | | | | Electronics | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rating Prediction | | | | Product Ranking | | | Rating Prediction | | | | Product Ranking | | |
| | | MSE | MAE | F-stat | p-value | AUC | NDCG | KL | MSE | MAE | F-stat | p-value | AUC | NDCG | KL |
| (a) | itemCF | 1.398 | <u>0.841</u> | 2.568 | 0.053 | 0.601 | 0.121 | 0.557 | <u>1.529</u> | <u>0.966</u> | 5.099 | <0.001 | 0.619 | 0.098 | 0.009 |
| | userCF | 1.880 | 0.946 | 3.889 | 0.009 | 0.504 | 0.123 | 0.303 | 2.487 | 0.980 | <u>1.501</u> | 0.186 | 0.503 | 0.087 | 0.009 |
| | PoissonMF | <u>1.168</u> | 0.859 | 9.600 | <0.001 | 0.638 | 0.151 | <u>0.001</u> | 1.628 | 1.035 | 4.112 | 0.001 | 0.565 | 0.085 | 0.014 |
| | MF | 1.176 | 0.859 | 9.805 | <0.001 | 0.817 | 0.179 | 0.015 | 1.590 | 1.025 | 3.447 | 0.004 | 0.591 | 0.091 | 0.012 |
| (b) | MF (reweighted) | 1.290 | 0.872 | 8.402 | <0.001 | <u>0.852</u> | <u>0.183</u> | 0.012 | 1.615 | 1.017 | 2.769 | 0.017 | 0.594 | 0.092 | <u>0.001</u> |
| | MF (corr.value) | 1.208 | 0.875 | 9.887 | <0.001 | 0.549 | 0.123 | 0.484 | 1.617 | 1.043 | 4.543 | <0.001 | 0.502 | 0.086 | 0.012 |
| | MF (corr.error) | 1.204 | 0.873 | <u>1.667</u> | 0.172 | 0.818 | 0.179 | 0.003 | 1.543 | 1.011 | 1.896 | 0.091 | <u>0.766</u> | <u>0.122</u> | 0.002 |

**Table 6: Recommendation results on *ModCloth* and *Electronics*. For rating predictions, MSE,MAE are used to evaluate the prediction accuracy while the F-statsitic (Eq. (2)) is used to evaluate the prediction fairness and its associated p-value is provided; for product rankings, AUC and NDCG are used to evaluate the recommendation accuracy while the KL-divergence (Eq. (3)) is used to evaluate the recommendation fairness. The most accurate and the fairest results are <u>underlined</u>.**

user-centric randomized experiments or natural experiments. In this way, causal conclusions and insights can be provided to product sellers and recommender system practitioners.

- **Algorithms.** Although we only focus on algorithms trained on explicit feedback, it is relatively intuitive to extend the proposed error correlation framework to other pointwise recommendation algorithms. Another direction is to address the marketing bias in pairwise ranking recommendation algorithms, where market fairness metrics and debiasing methods can be further explored to accommodate real-world scenarios.

## REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *RecSys*.
[2] Yoshua Bengio and Yann LeCun (Eds.). 2015. *ICLR*.
[3] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *KDD*.
[4] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. 2019. Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. In *AIES*.
[5] Al E Birdwell. 1968. A study of the influence of image congruence on consumer choice. *The Journal of Business* 41, 1 (1968), 76–88.
[6] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on Fairness, Accountability and Transparency*.
[7] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *RecSys*.
[8] Brian S Everitt. 1992. *The analysis of contingency tables*. Chapman and Hall/CRC.
[9] Prem Gopalan, Jake M Hofman, and David M Blei. 2015. Scalable Recommendation with Hierarchical Poisson Factorization.. In *UAI*.
[10] Stacy Landreth Grau and Yorgos C Zotos. 2016. Gender stereotypes in advertising: a review of current research. *International Journal of Advertising* 35, 5 (2016), 761–770.
[11] Edward L Grubb and Harrison L Grathwohl. 1967. Consumer self-concept, symbolism and market behavior: A theoretical approach. *Journal of Marketing* 31, 4 (1967), 22–27.
[12] Edward L Grubb and Gregg Hupp. 1968. Perception of self, generalized stereotypes, and brand selection. *Journal of Marketing research* 5, 1 (1968), 58–63.
[13] Jonathan L Herlocker, Joseph A Konstan, Al Borchers, and John Riedl. 1999. An algorithmic framework for performing collaborative filtering. In *SIGIR*.
[14] 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 (2015), 427–491.
[15] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *WSDM*.
[16] David G Kleinbaum, Lawrence L Kupper, Keith E Muller, and Azhar Nizam. 1988. *Applied regression analysis and other multivariable methods*. Vol. 601. Duxbury Press Belmont, CA.
[17] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.
[18] Frank Kressmann, M Joseph Sirgy, Andreas Herrmann, Frank Huber, Stephanie Huber, and Dong-Jin Lee. 2006. Direct and indirect effects of self-image congruence on brand loyalty. *Journal of Business research* 59, 9 (2006), 955–964.
[19] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.
[20] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 1 (2003), 76–80.
[21] Naresh K Malhotra. 1988. Self concept and product choice: An integrated perspective. *Journal of Economic Psychology* 9, 1 (1988), 1–28.
[22] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *CIKM*.
[23] Rishabh Misra, Mengting Wan, and Julian McAuley. 2018. Decomposing fit semantics for product size recommendation in metric spaces. In *RecSys*.
[24] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *EMNLP*.
[25] Badrul Munir Sarwar, George Karypis, Joseph A Konstan, John Riedl, et al. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW*.
[26] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *ICML*.
[27] M Joseph Sirgy. 1982. Self-concept in consumer behavior: A critical review. *Journal of consumer research* 9, 3 (1982), 287–300.
[28] M Joseph Sirgy, Dhruv Grewal, Tamara F Mangleburg, Jae-ok Park, Kye-Sung Chon, C Berkman Claiborne, Jotindar S Johar, and Harold Berkman. 1997. Assessing the predictive validity of two methods of measuring self-image congruence. *Journal of the academy of marketing science* 25, 3 (1997), 229.
[29] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *RecSys*.
[30] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *NeurIPS*.
[31] Xiaoying Zhang, Junzhou Zhao, and John Lui. 2017. Modeling the assimilation-contrast effects in online product rating systems: Debiasing and recommendations. In *RecSys*.
[32] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *CIKM*.