Modeling the Assimilation-Contrast Effects in Online Product Rating Systems: Debiasing and Recommendations

Xiaoying Zhang
The Chinese University of Hong Kong
xyzhang@cse.cuhk.edu.hk

Junzhou Zhao
The Chinese University of Hong Kong
junzhouzhao@gmail.com

John C.S. Lui
The Chinese University of Hong Kong
cslui@cse.cuhk.edu.hk

ABSTRACT

The *unbiasedness* of online product ratings, an important property to ensure that users' ratings indeed reflect their true evaluations to products, is vital both in shaping consumer purchase decisions and providing reliable recommendations. Recent experimental studies showed that distortions from historical ratings would ruin the *unbiasedness* of subsequent ratings. How to "*discover*" the distortions from historical ratings in each *single* rating (or at the micro-level), and perform the "*debiasing operations*" in real rating systems are the main objectives of this work.

Using 42 million real customer ratings, we first show that users either "*assimilate*" or "*contrast*" to historical ratings under different scenarios: users conform to historical ratings if historical ratings are not far from the product quality (assimilation), while users deviate from historical ratings if historical ratings are significantly different from the product quality (contrast). This phenomenon can be explained by the well-known psychological argument: the *"Assimilate-Contrast"* theory. However, none of the existing works on modeling historical ratings' influence have taken this into account, and this motivates us to propose the *Historical Influence Aware Latent Factor Model* (HIALF), the first model for real rating systems to capture and mitigate historical distortions in each *single* rating. HIALF also allows us to study the influence patterns of historical ratings from a modeling perspective, and it perfectly matches the assimilation and contrast effects we previously observed. Also, HIALF achieves significant improvements in predicting subsequent ratings, and accurately predicts the relationships revealed in previous empirical measurements on real ratings. Finally, we show that HIALF can contribute to better recommendations by decoupling users' real preference from distorted ratings, and reveal the intrinsic product quality for wiser consumer purchase decisions.

1 INTRODUCTION

Online rating system is perhaps one of the most important modules in a wide variety of contemporary web applications ranging from e-commerce websites [18, 27] to online video/news platforms [7, 31]. Such online rating systems allow users to rate items (e.g., products, videos, etc.) they have recently consumed, and these ratings can help subsequent users in making decisions on whether to consume this item or not. In order to have correct subsequent decision making, the *unbiasedness* of ratings, a property to ensure that users' ratings indeed reflect their true evaluations to the product, is crucial. Furthermore, unbiased users' ratings are also important to recommender systems so that they can provide reliable recommendations.

However, recent experimental studies [3, 19, 22, 30] showed that the disclosed historical ratings would ruin the *unbiasedness* of subsequent ratings, making them inaccurate to convey users' intrinsic evaluations to products. Such distortions bring both *macro-level* and *micro-level* effects. At the macro level, the distortions from historical ratings will make overall rating distribution deviate from the intrinsic product quality, thereby misleading subsequent consumers to wrong purchase decisions [19, 22, 30]. At the micro level (or at the granularity of each single rating), the distortion in the rating provides an adulterated view of user's preference for the product, weakening recommender systems' ability to provide high-quality recommendations [3]. As in [3], even for products with the same quality, users tend to rate higher when they observe high historical ratings as compared to low historical ratings. Thus, when a user rates a product high under high historical ratings, the high rating may not suggest the user's high preference to the product anymore, since it may be the result of high historical ratings.

Recently, Wang et al. [28] studied the macro-level influence from historical ratings. However, to debias the historical distortions in recommendations, we need a *micro-level* model to characterize the historical ratings' influence in each *single* rating. Previously, several works [2, 13] tried to mitigate the micro-level historical ratings' influence with an assumption that we know *users' intrinsic ratings*, the ratings given when users couldn't observe historical ratings. However, their models are inapplicable in real rating systems where users' intrinsic ratings are usually latent. To the best of our knowledge, there is no work to characterize and debias the micro-level influence from historical ratings in real rating systems. The main challenge is that people do not fully understand how historical ratings affect the user who gives the next rating.

Present work. The goal of this work is to develop a model for *real rating systems* to accurately characterize and debias the influence from historical ratings in each *single* rating microscopically. To handle the challenge mentioned before, we analyze real ratings to understand how historical ratings affect user's rating behavior.

In this work, we first analyze a dataset of 42 million ratings from *Tripadvisor* and *Amazon*. We find that users either *assimilate* or *contrast* to historical ratings under different scenarios: a user tends to give a rating similar to historical ratings when historical ratings are not far from the product quality (assimilation), while deviating from historical ratings when historical ratings differ significantly from the product quality (contrast). In fact, this phenomenon can be well explained by the *"Assimilate-Contrast"* theory [4] in psychology.

RecSys '17, August 27-31, 2017, Como, Italy

ACM ISBN 978-1-4503-4652-8/17/08…$15.00
https://doi.org/10.1145/3109859.3109885

Then, we find that the previous works [2, 13, 28] were unable to explain our empirical results. Thus we propose *Historical Influence Aware Latent Factor Model* (HIALF), the first model designed for *real rating systems* to capture and mitigate the *micro-level* influence from historical ratings. In HIALF, we do not make any assumptions about influence patterns of historical ratings, but we discover the most likely influence pattern from data. The discovered influence patterns via HIALF match perfectly with the assimilation and contrast effects in empirical observations. Compared with previous methods, HIALF reveals the closest fitting to the relationships observed in previous empirical measurements on real ratings, and significantly reduces the mean-squared error (MSE) in predicting subsequent ratings, i.e., up to 39% as compared with HEARD [28] and 12% as compared with the standard latent factor model [21].

Finally, we demonstrate two applications of HIALF. HIALF enables us to separate users' intrinsic interests from historical distortions, leading to better product recommendations. Also, we can directly compare products by their intrinsic qualities, without being misguided by distorted historical ratings.

Contributions. Overall, we make the following contributions:

- **Observations.** We first reveal the assimilation and contrast effects in user's rating behavior caused by historical ratings. We also provide an explanation for our observations by a well-known psychological theory (Section 2).
- **Modeling.** We develop the first model (HIALF) for real rating systems to characterize and mitigate historical distortions in each single rating microscopically (Section 3).
- **Performance.** The discovered influence patterns of historical ratings via HIALF perfectly match the assimilation and contrast effects in observations. Moreover, HIALF achieves significantly improvements in predicting subsequent ratings, and accurately fits the relationships revealed in empirical measurements on real ratings (Section 4).
- **Applications.** HIALF can contribute to better recommendations by separating users' intrinsic interests from historical distortions. It can also facilitate wiser purchase decisions by revealing the intrinsic product quality (Section 5).

2 HOW HISTORICAL RATINGS AFFECT THE NEXT SINGLE RATING

We conduct empirical measurements on real world datasets to study how historical ratings affect its next single rating. In this section, we first describe these rating datasets, then we discuss how to measure the impact of historical ratings on its next rating. Finally, we propose an explanation of our empirical observations, and verify that the existing works [2, 13, 28] cannot explain our observations, which motivates us to design a model for real rating systems to depict the micro-level historical ratings' influence in the next section.

2.1 Rating Datasets

We first introduce two large scale public available rating datasets from Amazon[1] and TripAdvisor[2], respectively.

Amazon is a popular e-commerce website that allows users to review and rate products they recently consumed, e.g., books, clothes, etc. In the Amazon dataset [18], we focus on ratings of the top four largest categories: books, movies, electronics, and clothes. These four categories cover about 48.8% of all products, and 50.4% of all ratings on Amazon. The dataset spans from May 1996 to July 2014.

TripAdvisor is a popular travel website that provides reviews and ratings of travel-related contents, e.g., hotels, restaurants, etc. We use the entire ratings on it from April 2001 to September 2012 [26].

Table 1 summarizes the basic statistics of our dataset.

Table 1: Summary of rating datasets.

category	# products	# users	# ratings
Amazon-books	2, 370, 585	8, 026, 324	22, 507, 155
Amazon-clothes	1, 503, 384	3, 117, 268	5, 748, 920
Amazon-electronics	498, 196	4, 261, 096	7, 824, 482
Amazon-movies	208, 321	2, 088, 620	4, 607, 047
TripAdvisor	12, 730	781, 329	1, 621, 956

2.2 Empirical Measurements and Observations

We conduct empirical measurements on the above datasets to study how historical ratings affect its next rating.

Let $r_{p,i}$ denote the i-th rating of product p, and let $\mathcal{H}_{p,i} \triangleq (r_{p,1}, \ldots, r_{p,i-1})$ denote a sequence of $i-1$ ratings of product p received before $r_{p,i}$ (in the chronological order of receiving time). $\mathcal{H}_{p,i}$ will be referred to as the *historical ratings* of $r_{p,i}$.

We want to measure how historical ratings $\mathcal{H}_{p,i}$ affect its next rating $r_{p,i}$. Intuitively, there are two factors that could affect a user's decision on rating a product: (1) the product quality; (2) other users' ratings/reviews to which the user was exposed.

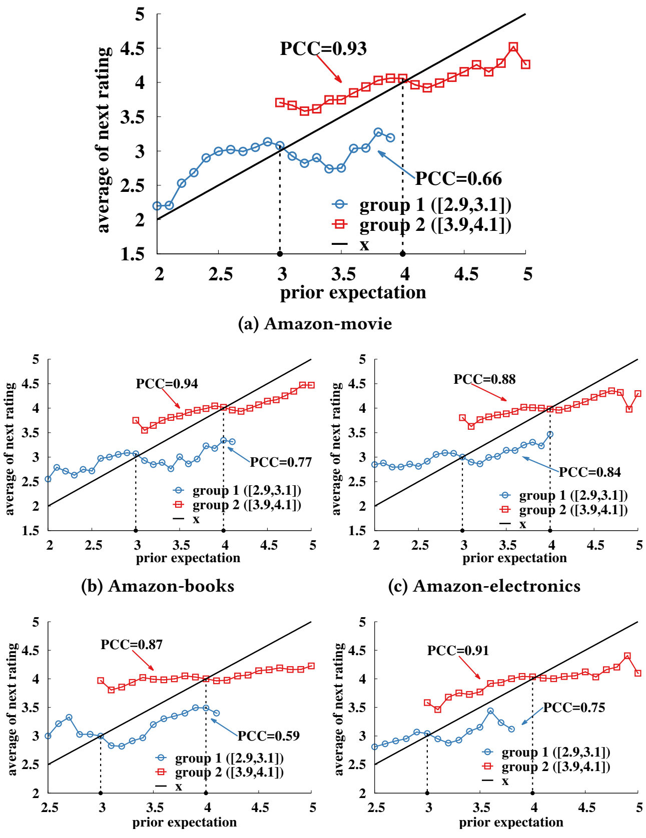
The first factor is latent and around the average of ratings given by a large population who were not exposed to historical ratings [25]. To process our dataset, we group products with similar average ratings into one group such that each group has a maximum deviation of 0.2 in the average rating. For example, consider the two selected groups of products with average ratings in $[2.9, 3.1]$ and $[3.9, 4.1]$, and we assume the first (second) group has an *approximately true quality* of 3 (4). We only consider groups containing more than 100 products and on average, each dataset has 10 groups.

Then, in each product group, for each rating $r_{p,i}$, we first calculate its *prior expectation* formed on historical ratings $\mathcal{H}_{p,i}$: $e_{p,i} = \frac{1}{i-1}\sum_{k=1}^{i-1} r_{p,k}$, resulting in a pair $(e_{p,i}, r_{p,i})$[3]. We find that a set of pairs $\{(\dot{e}, \dot{r}_1), \ldots, (\dot{e}, \dot{r}_k)\}$ have the same prior expectation $\dot{e}$, but the ratings $\{\dot{r}_1, \ldots, \dot{r}_k\}$ are given by thousands of different users. We aggregate ratings $\{\dot{r}_1, \ldots, \dot{r}_k\}$ to get $\bar{r} = \frac{1}{k}\sum_{i=1}^{k} \dot{r}_i$. Thus, the resulting list $\{(\dot{e}, \bar{r})\}$ describes how prior expectation $\dot{e}$ affects its next rating $\bar{r}$, on average, in this product group. Finally, we plot the relationship between prior expectation $\dot{e}$ and the average of the next rating $\bar{r}$ for each selected product group in Figure 1. Relationships in other groups are similar with the two selected groups.

Some may argue that the user's personal taste is another factor that affects the user's rating. However, the user's personal taste will not affect the plotted relationships, because for each prior expectation $\dot{e}$, we aggregate thousands of users' ratings to obtain the average of the next rating $\bar{r}$, the aggregated distortions of thousands of users' personal tastes will be insignificant and can be ignored.

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

[3] In all our analysis, we round $e_{p,i}$ to one decimal place.



(a) Amazon-movie

(b) Amazon-books (c) Amazon-electronics

(d) Amazon-clothes (e) Tripadvisor

Figure 1: Relationship between prior expectation $\dot{e}$ and the average of the next rating $\bar{r}$. The group 1 (group 2) contains products with average ratings in [2.9, 3.1] ([3.9, 4.1]). We observe that products' historical ratings do affect its next rating, and each curve with □ (○) is divided into two parts by the group's *approximately true quality* (3 for group 1, and 4 for group 2).

Observations. We obtain two main observations from Figure 1:

- Products' historical ratings do affect the next rating. For example, in Figure (1a), the Pearson correlation coefficient of group 1 (products with average ratings in $[2.9, 3.1]$) and group 2 (products with average ratings in $[3.9, 4.1]$) are 0.66 and 0.93 respectively. In general, the Pearson correlation coefficients (PCC) are in the range $[0.59, 0.94]$, this reflects a positive correlation between prior expectation and the next rating.
- Each curve with □ (○) is divided into two parts by the group's *approximately true quality* (3 for group 1, and 4 for group 2). The black line represents a hypothetical linear relationship between prior expectation and its next rating, i.e., the user will give a 4-star rating as long as his prior expectation is 4. Take the group 2 in Figure (1a) as an example, when prior expectation is below the group's approximately true quality of 4, it will receive a rating higher than the prior expectation, on average; and when prior expectation is above the group's approximately true quality of 4, it will receive a rating lower than the prior expectation, on average. It is important to note that this phenomenon is *consistent* among all groups of products in our dataset, and it is interesting to find an explanation of this result.

2.3 Proposed Explanation of Observations

Let us now answer two fundamental questions: (1) why do historical ratings influence its next rating? (2) why does the influence of historical ratings behave consistently like those in Figure 1?

One possible answer to the first question is that different historical ratings lead the user to form different prior expectations for the product, which impact the user's overall satisfaction with the product (the given rating). Before consuming a product, a customer usually refers to previous aggregated ratings to see whether the product really meets his needs. At this stage, he forms his "prior expectation" for that product. Using the customer satisfaction theory [20], user's prior expectation of the product and the product quality together determine the user's satisfaction on the product. Thus, different historical ratings lead to different prior expectations, which in turn affect the next single rating.

To answer the second question, we refer to three well-known psychological theories [4] which describe how the user's prior expectation for the product and the product quality together determine the user's overall satisfaction with the product. Figure (2a) shows the sample representations of three theories. The product quality is 3 and is represented by the line parallel with x axis.

- **"Assimilate" theory**: The user's satisfaction of the product is always similar to his prior expectation (the orange line with △).
- **"Contrast" theory**: The customer will magnify the difference between his prior expectation for the product and the product quality; i.e., if his prior expectation is below (above) the product quality, the user will evaluate the product more (less) favorably than the product quality (the purple line with ○).
- **"Assimilate-Contrast" theory**: If the disparity between his prior expectation and the product quality can be accepted by the user (in $[\theta, \sigma]$ in Figure (2a)), the user's satisfaction with the product assimilates to his prior expectation; otherwise, the difference between the prior expectation and the product quality tends to be magnified (the red line with □).

One interesting question is which theory can explain our empirical observations. To answer this question, we combine all groups with the same average rating range in five datasets. For example, Figure (2b) illustrates the relationship for all products with average ratings in $[2.9, 3.1]$. We choose this group because its approximately true quality is in the middle of the rating scale of *Amazon* and *Tripadvisor*. One can see that the relationship follows the *"Assimilate-Contrast"* theory in that (1) prior expectation has a positive correlation with its next rating (contradicting with the *"Contrast"* theory), and (2) when prior expectation is in $[1, 1.8)$ and $(4.6, 5]$, the average of the next rating diverges from the increasing trend in $[1.8, 4.6]$ (contradicting with the *"Assimilate"* theory).

Hypothesis Test. Let us now design a statistical hypothesis test to see whether the *"Assimilate-Contrast"* theory can explain our observations in general. The *"Assimilate-Contrast"* theory differs from the other two theories because its slope changes when the difference between prior expectation and the product quality is large, for example, when prior expectation is in $[1, \theta)$ and $(\sigma, 5]$ in Figure (2a). In our test, each data point is a pair (e, r_e), where e is the prior expectation, and r_e is the rating given under e. Thus, we can construct four sets. The set $C_1 (C_3)$ contains all data points

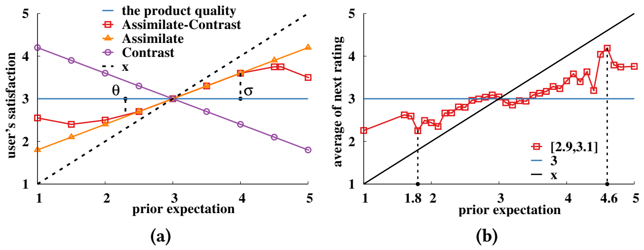


Figure 2: (a) Sample representations of three theories. (b) Relationship between prior expectation and the average of the next rating in products with average ratings in $[2.9, 3.1]$ **in all five datasets.**

with prior expectation e much smaller (larger) than the product quality q, $C_1 = \{(e, r_e)|e < \theta < q\}$ ($C_3 = \{(e, r_e)|e > \sigma > q\}$). The set $C_2(C_4)$ contains data points with prior expectation e smaller (larger) than but near the product quality q, $C_2 = \{(e, r_e)|\theta < e < q\}$ ($C_4 = \{(e, r_e)|q < e < \sigma\}$). Then, we calculate the slope between $(\theta, \bar{r}_\theta)$ and each point in $C_1(C_2)$, putting it in $S_1(S_2)$, i.e., $S_i = \{\frac{r_e - \bar{r}_\theta}{e - \theta}|(e, r_e) \in C_i\}, i = 1, 2$. Here, $\bar{r}_\theta$ is the average of ratings under prior expectation θ. Similarly, we calculate the slope between $(\sigma, \bar{r}_\sigma)$ and each point in $C_3(C_4)$, getting $S_3(S_4)$. We will discuss how we set θ and σ in the next paragraph. If the *"Assimilate-Contrast"* theory holds, the mean of $S_1(S_3)$ should be smaller than the mean of $S_2(S_4)$. Otherwise, the mean of $S_1(S_3)$ equals to the mean of $S_2(S_4)$. We use student t-test to examine whether there exists a significant difference between the mean of $S_1(S_3)$ and $S_2(S_4)$.

Specifically, we first group products with similar average ratings as before. In each group, we divide the $\{(e, r_e)\}$ pairs into distinct partitions such that pairs in each partition share a unique prior expectation. We discretize real-valued e by equal-interval partition, i.e., prior expectations in $[a - \epsilon, a + \epsilon]$[4] are taken as the same prior expectation as a. Then, we get $S_i, i = 1, 2, 3, 4$ and apply the student t-test [5]. We use Welch-Satterthwaite approximation [23] to get the degrees of freedom. Let $\hat{\theta}$ ($\hat{\sigma}$) be the fourth smallest (largest) prior expectations, then we set $\theta = \min\{\hat{\theta}, q - 1\}$, and $\sigma = \max\{\hat{\sigma}, q + 1\}$. The null hypothesis in our test is that there exists no statistically significant difference between the mean of $S_1(S_3)$ and $S_2(S_4)$, while the alternate hypothesis is that the mean of $S_1(S_3)$ is smaller than the mean of $S_2(S_4)$. We test all hypothesis at the 0.05 significance level. We observe that 11 out of 12 groups reject the null hypothesis of the t-test between S_1 and S_2, and 7 out of 8 groups reject the null hypothesis of the t-test between S_3 and S_4. Since almost all groups reject the null hypothesis, the *"Assimilate-Contrast"* theory is a more appropriate theory to explain how customers are affected by historical ratings.

2.4 Deficiencies of existing works

Next, we show that existing works on modeling historical ratings' influence [2, 13, 28] fail to explain our previous observations.

In our dataset, users' intrinsic ratings are latent, thus we are unable to build the model in [2, 13]. Wang et al. [28] developed HEARD to model how historical ratings $\mathcal{H}_{p,i}$ influence the general rating distribution after next M ratings $x_{p,i+M}$ at the macro level. Here, for a one-to-K star rating system, $x_{p,i+M} \triangleq [x^{(1)}_{p,i+M}, \ldots, x^{(K)}_{p,i+M}]$, where $x^{(k)}_{p,i+M}$ represents the proportion of level-k ratings in the first $(i + M - 1)$ ratings of the product p. Note that the goal of HEARD is fundamentally different from ours. However, given historical ratings $\mathcal{H}_{p,i}$, the probabilistic model of HEARD can reveal the probability $P(r_{p,i} = k|\mathcal{H}_{p,i}), \forall k \in \{1, \ldots, K\}$. Hence, HEARD can be taken as a model to predict the next rating $r_{p,i}$ given its history $\mathcal{H}_{p,i}$. Thus, we perform experiments to see whether HEARD can reveal our previous observations. Specifically, we first train HEARD with each dataset. Then, we select the same groups of products in each dataset as in Figure 1. In each selected group, for each rating, given its historical ratings $\mathcal{H}_{p,i}$, we use HEARD to predict the next rating $r^H_{p,i} = argmax_k P(r_{p,i} = k|\mathcal{H}_{p,i})$. We also calculate its prior expectation $e_{p,i} = \frac{1}{i-1}\sum_{k=1}^{i-1} r_{p,k}$ based on real ratings, obtaining a pair $(e_{p,i}, r^H_{p,i})$. Finally, for each distinct prior expectation e, we calculate the average of the next HEARD-generated ratings under e, denoting as r^H_e.

Before checking the slope changes as in hypothesis tests, we first check whether the resulting list $\{(e, r^H_e)\}$ meets $r^H_e \le e$, when $e \ge q$, as in Figure 1. Here q refers to the approximately true quality of the product group. Let E^+_q denote those prior expectations larger than q: $E^+_q = \{e|e - q \ge 0\}$. We calculate the average deviation from e to r^H_e when $e \ge q$: $d^H = \frac{1}{|E^+_q|}\sum_{e \in E^+_q}(e - r^H_e)$. Let r^*_e be the average of the next *real* rating given under prior expectation e. We also calculate the average deviation from e to r^*_e when $e \ge q$, which we denote as d^*. Note that d^* is always positive in the real rating datasets, because $r_e \le e$, when $e \ge q$ in Figure 1. We present d^H and d^* in both groups on all five datasets in Table 2. From Table 2, we observe that all d^H are negative and significantly different from the positive d^*. For example, in *Amazon-books*, for all e, on average, HEARD predicts a larger r^H_e than e in both groups ($d^H < 0$), while in real ratings, r_e should be smaller than e ($d^* > 0$). This already suggests that HEARD *fails* to explain our observations in real rating datasets, and there is no need for the hypothesis tests. The deficiency of HEARD is because HEARD mainly focuses on the macro-level historical ratings' influence in overall rating distribution, rather the micro-level historical influence in each singe rating.

Table 2: d^H and d^* on all five datasets.

category	HEARD (d^H)		real ratings (d^*)	
	group 1	group 2	group 1	group 2
Amazon-books	-0.7899	-0.4305	0.5697	0.3681
Amazon-clothes	-0.8629	-0.4438	0.3733	0.4396
Amazon-electronics	-0.4944	-0.4421	0.4202	0.3868
Amazon-movies	-0.8194	-0.4332	0.5333	0.3913
TripAdvisor	-0.5097	-0.2339	0.2784	0.4368

3 PROPOSED MODEL

In this section, we describe in detail the *Historical Influence Aware Latent Factor Model* (HIALF) which leverages previous observations to characterize the micro-level influence from historical ratings in real rating systems. Our objectives are: (1) to model the influence of historical ratings so as to do a better prediction of the next rating; (2) to reveal the intrinsic qualities of products and users' intrinsic preference so to do a better job in product recommendations.

[4] we set $\epsilon = 0.1$.

[5] We discard groups containing less than 100 products and prior expectations followed by less than 100 ratings due to their low statistical reliability.

3.1 Preliminary: Latent Factor Model

One can first consider using the classical latent factor (LF) model [21] to predict the rating $r_{u,p}$ for user u and product p as:

$$r_{u,p} = g + b_u + b_p + \boldsymbol{x_u}^T \boldsymbol{y_p} \tag{1}$$

Here g is the overall rating for an arbitrary user and product; b_u and b_p denote the user and item bias, respectively; $\boldsymbol{x_u}$ and $\boldsymbol{y_p}$ represent vectors of latent features for user u and product p.

We need to emphasize that the standard latent factor model cannot explain our empirical observations because it does not consider factors due to the effects of historical ratings to subsequent ratings. We mention the standard latent factor model because HIALF is an enhanced version of the LF model.

3.2 Historical Influence Aware Latent Factor Model (HIALF)

Let the term *user u's experienced quality of product p* refer to product p's quality in user u's view. We use $h_{p,i}$ to represent the distortion from historical ratings $\mathcal{H}_{p,i}$.

In HIALF, the i-th rating of product p given by user u is mainly taken as a combination of two factors: (1) user u's experienced quality of product p, denoted as $q_{u,p}$; (2) the distortion from historical ratings $h_{p,i}$. The first factor is determined by product p's intrinsic quality and user u's overall interest in product p. We model it by

$$q_{u,p} = g + b_p + \boldsymbol{x_u}^T \boldsymbol{y_p} \tag{2}$$

According to previous observations, the second factor $h_{p,i}$ depends on the discrepancy between $q_{u,p}$ and the prior expectation formed on the historical ratings (i.e., $e_{p,i}$). Thus, we use a categorical function $\boldsymbol{\beta}(x)$ to represent the induced bias when the difference between $e_{p,i}$ and $q_{u,p}$ is x, i.e., $x = e_{p,i} - q_{u,p}$. We call $\boldsymbol{\beta}(x)$ as *the disconfirmation bias curve*. Moreover, applying Latané's theory [14], the size of historical ratings $|\mathcal{H}_{p,i}|$ will boost the distortion $h_{p,i}$. For example, 100 historical ratings will exert a larger influence on the next rating than only 1 historical rating. Thus, let $f(x)$ be a scaling function to represent the magnitude of impact by historical ratings of size x. We have:

$$h_{p,i} = f(|\mathcal{H}_{p,i}|)\boldsymbol{\beta}(e_{p,i} - q_{u,p}) \tag{3}$$

All in all, HIALF predicts $\hat{r}_{p,i,u}$ for the i-th rating of product p given by user u as follows:

$$\begin{aligned} \hat{r}_{p,i,u} &= b_u + q_{u,p} + \alpha_u h_{p,i} \\ &= g + b_u + b_p + \boldsymbol{x_u}^T \boldsymbol{y_p} + \alpha_u f(|\mathcal{H}_{p,i}|)\boldsymbol{\beta}(e_{p,i} - q_{u,p}) \end{aligned} \tag{4}$$

Here, $g, b_u, b_p, \boldsymbol{x_u}, \boldsymbol{y_p}$ take on the same roles as in the basic latent factor model; α_u models how easily user u will be influenced by historical ratings. A larger α_u means that user u is easier to be affected. Next, we describe how to model $\boldsymbol{\beta}(x)$, $f(x)$, and give a more realistic formula of $e_{p,i}$.

Modeling the disconfirmation bias curve $\boldsymbol{\beta}(x)$. We use a data-driven approach to model $\boldsymbol{\beta}(x)$, i.e., we do not constrain the form of $\boldsymbol{\beta}(x)$ (i.e., to be linear or quadratic). Instead, we learn the most appropriate format from data. We expect the learned $\boldsymbol{\beta}(x)$ can match the assimilation and contrast effects in previous observations.

Online rating systems usually have a limited rating range. For example, *Amazon* and *Tripadvisor* adopt one-to-five-star rating system. Thus, $x = e_{p,i} - q_{u,p}$ is in a fixed known range $[x_a, x_b]$. For example, on both *Amazon* and *Tripadvisor*, $x \in [-4, 4]$ since both $e_{p,i}$ and $q_{u,p}$ are in $[1, 5]$. In this work, we use non-parametric kernel regression [29] to model $\boldsymbol{\beta}(x)$.

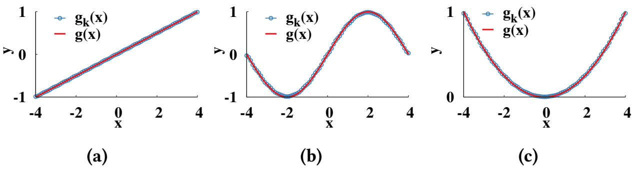


Figure 3: Using kernel function to approximate three different functions with x in $[-4, 4]$: (a) $g(x) = \frac{x}{4}$; (b) $g(x) = \sin(\frac{\pi x}{4})$; (c) $g(x) = (\frac{x}{4})^2$. In kernel function, we use $\{x_1, \ldots, x_n\} = \{-4, -3.5, -3, \ldots, 3, 3.5, 4\}$ and $\kappa = 10$.

In kernel regression, given a set of i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ from model $y_i = g(x_i) + \epsilon_i$, where ϵ_i represents the noise from the standard normal distribution, we can approximate $g(x)$ by a kernel function $g_k(x) = \frac{\sum_{i=1}^n w(x,x_i)\cdot y_i}{\sum_{i=1}^n w(x,x_i)}$. The term $w(x, x_i)$ gives a greater weight to x that is closer to x_i, and we select $w(x, x_i) = exp(-\kappa(x - x_i)^2)$, where κ controls the smoothness of the function. Figure 3 shows examples using kernel methods to approximate three different $g(x)$ with x in $[-4, 4]$, and $g_k(x)$ always gives a good approximation to $g(x)$.

Thus, if we can get a set of samples $\{(e_l, v_l)\}_{l=1}^n$ from the disconfirmation bias curve, i.e., $v_l = \boldsymbol{\beta}(e_l) + \epsilon_l$, where ϵ_l represents the noise from the standard normal distribution, we represent $\boldsymbol{\beta}(x)$ as:

$$\boldsymbol{\beta}(\boldsymbol{x}) = \frac{\sum_{l=1}^n w(x, e_l) \cdot v_l}{\sum_{l=1}^n w(x, e_l)} \tag{5}$$

To obtain the set of samples $\{(e_l, v_l)\}_{l=1}^n$, we let $\{e_1, \ldots, e_n\}$ be uniformly distributed in the known range of x ($[x_a, x_b]$), i.e., in our dataset, we set $\{e_1, \ldots, e_n\} = \{-4, -3.5, \ldots, 3.5, 4\}$. And we take $\{v_1, \ldots, v_n\}$ as parameters and learn them from data.

Modeling magnifying curve $f(x)$. Intuitively, the more historical ratings exist, the larger the magnifying effect will be. The previous psychological study [14] showed that the slope of $f(x)$ decreases as x increases, but the slope remains positive. In this work, we use the following *magnitude function* $f(x)$ to describe the magnifying effect of historical ratings with a size x.

$$f(x) = \frac{a}{1 + exp(-b * x)} - \frac{a}{2} \tag{6}$$

The first component is a sigmoid function while the second component (subtracting $a/2$) is to ensure that $f(0) = 0$, because when we do not have any historical ratings, no magnifying effect exists.

Modeling prior expectation $e_{p,i}$. In previous measurements, we used the average of historical ratings as prior expectation $e_{p,i}$. In reality, users focus more on recent ratings instead of all ratings of a product. Hence, we represent $e_{p,i}$ by the following general formula:

$$e_{p,i} = \frac{\sum_{k=1}^{i-1} \xi(i-k) \cdot r_{p,k}}{\sum_{k=1}^{i-1} \xi(i-k)} \tag{7}$$

Here, $\xi(d) = exp(-\gamma * d)$ denotes an exponential triggering kernel, which models the decay of influence; $r_{p,k}$ is the k-th real rating of product p; γ controls the extent to which users prefer recent ratings. If γ is 0, then $e_{p,i}$ is exactly the average of historical ratings. A larger γ means that users focus more on recent ratings. In our case, γ is set by cross-validation.

3.3 Model Inference

Overall, our goal is to solve the following optimization problem:

$$\min_{\Theta} \sum_{(p,i,u)\in\mathcal{K}} (r_{p,i,u} - \hat{r}_{p,i,u})^2 + \lambda_{rec}(b_u^2 + b_p^2 + ||\boldsymbol{x_u}||_2^2 + ||\boldsymbol{y_p}||_2^2)$$
$$+ \lambda_f(a^2 + b^2) + \lambda_\beta(\sum_l v_l^2) + \lambda_\alpha(\alpha_u^2)$$

Here, $\Theta = \{g, \{b_u\}, \{b_p\}, \{\boldsymbol{x_u}\}, \{\boldsymbol{y_p}\}, \{\alpha_u\}, a, b, \{v_l\}\}$; $r_{p,i,u}$ is the real rating; $\hat{r}_{p,i,u}$ is the predicted rating by HIALF (Equation (4)); $\mathcal{K}$ contains all (p, i, u) pairs, and a pair (p, i, u) represents that the i-th rating of product p in the dataset is given by user u. Using this objective function, we aim to make $\hat{r}_{p,i,u}$ as close as possible to the real rating $r_{p,i,u}$. $\lambda_{rec}, \lambda_f, \lambda_\beta, \lambda_\alpha$ are regularization hyperparameters to prevent overfitting. We use stochastic gradient descent (SGD) algorithm to learn parameters, which is widely used in previous works [8, 11, 12], due to its efficiency.

4 EXPERIMENTS

We conduct experiments on real rating datasets (Table 1) to compare the performance of our model (HIALF) with state-of-the-art models. We compare different models by evaluating: (1) how accurate a model could predict the subsequent ratings, and (2) how well a model could fit the previous empirical observations in real ratings.

4.1 Validating The Disconfirmation Bias Curve

One important thing we need to check is whether the disconfirmation bias curve $\boldsymbol{\beta}(x)$ meets with the *"Assimilate-Contrast"* theory because this will dictate the accuracy of HIALF.

As in Figure 4, all learned $\boldsymbol{\beta}(x)$ perfectly match the *"Assimilate-Contrast"* theory. $\boldsymbol{\beta}(x)$ on *Amazon-books*, *Amazon-movie*, *Amazon-electronics* and *Tripadvisor*, have similar formats with the sample representation of the *"Assimilate-Contrast"* theory in Figure (2a). $\boldsymbol{\beta}(x)$ on *Amazon-clothes* also follows the theory: in the range $[0, 1]$, the bias roughly equals to difference between prior expectation and the product quality, while deviating it out of the range.

We also notice that $\boldsymbol{\beta}(x)$ is close to 0 for some x, for example, x in $[-4, -3]$ or $[3, 4]$ in Figure (4a). There are two possible reasons. For one thing, x seldom achieves values in these ranges. Take $x = -3$ as an example, it means that a user takes an inferior product in others' views (i.e., forming a prior expectation as 1-star) as a good 4-star product. In reality, such large discrepancy rarely occurs. With a high probability, an inferior product in many users' view is truly a bad product. Then with constraints on value of v_l (i.e., λ_β), $\boldsymbol{\beta}(x)$ in the above ranges is close to 0. For another, from a psychological point of view, as mentioned in [32], if users find others' opinions highly contradict with their own opinions, they may tend to insist on their own opinions.

4.2 Predicting Subsequent Ratings

For the rating sequence of each product, we split it into the training subsequence and the testing subsequence, and put the two subsequences into the training set and the testing set, respectively. We train the model on the training set, and validate the model on the testing set in terms of mean squared error (MSE). We compare HIALF with several state-of-the-art models: HEARD [28], latent factor (LF) model [21], and also a variant of HIALF model, denoted by HIALF-AVG. In HIALF-AVG, prior expectation is taken as the average of historical ratings without emphasis on recent ratings.

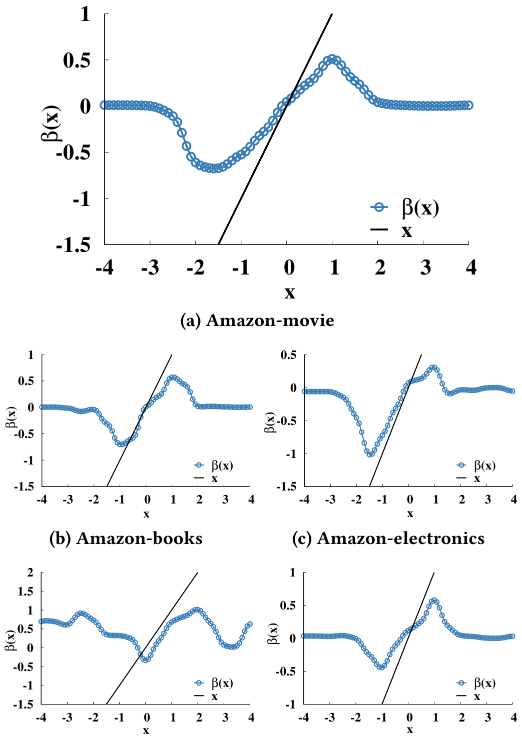


Figure 4: The learned disconfirmation bias curve $\boldsymbol{\beta}(x)$. All $\boldsymbol{\beta}(x)$ perfectly match the *"Assimilate-Contrast"* theory.

Results. Table 3 shows that our model significantly outperforms alternatives on all datasets. On average, HIALF achieves a 33% reduction in MSE compared to HEARD, and a 6% reduction to LF.

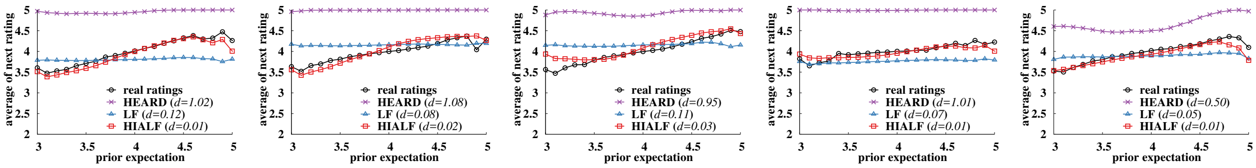
Note that improvements over LF are bounded by the extent to which ratings in the dataset are affected by historical ratings. We apply curve fitting to all lines in Figure 1 with a linear model $y = ax + b$. Here, the slope a represents the average increase in the next rating when prior expectation moves from x to $x + 1$, thus a larger a implies the larger distortions from prior expectations (historical ratings). The average slope of two fitting curves in *Amazon-movie* is the largest (0.34), while *Amazon-clothes* reveals the smallest average slope as 0.22. In other words, *Amazon-movie* suffers the largest distortions from historical ratings, while *Amazon-clothes* suffers the smallest historical distortions. From Table 3, HIALF has the most significant benefits on *Amazon-movie* over LF and the least benefits on *Amazon-clothes* over LF. Furthermore, HIALF is consistently more accurate than HIALF-AVG, because users focus more on recent ratings when shaping prior expectations.

4.3 Fitting Empirical Observations

Next, we re-do the empirical measurements in Section 2 with the predicted ratings by HEARD, LF, HIALF, respectively. Note that an accurate model should reveal a similar relationship as in our previous observations in real ratings.

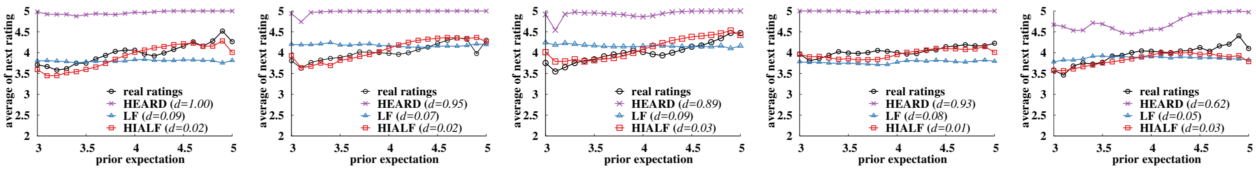
Table 3: MSE on five datasets

	Amazon-movie	Amazon-books	Amazon-electronics	Amazon-clothes	Tripadvisor
HEARD	1.5826	1.5548	3.1170	2.1550	1.3135
LF	1.2794	1.0777	1.9634	1.4123	1.0074
HIALF-AVG	1.2054	1.0619	1.9357	1.3985	0.9805
HIALF	**1.1194**	**1.0318**	**1.8764**	**1.3759**	**0.9405**
benefits of HIALF over HEARD	29.27%	32.83%	39.80%	35.17%	28.40%
benefits of HIALF over LF	12.51%	4.26%	4.43%	2.58%	6.64 %



(a) Amazon-movie (b) Amazon-electronics (c) Amazon-books (d) Amazon-clothes (e) Tripadvisor

Figure 5: Relationship between prior expectation (defined in Equation (7)) and the average of the next rating. A smaller d implies a better fitting.



(a) Amzon-movie (b) Amazon-electronics (c) Amazon-books (d) Amazon-clothes (e) Tripadvisor

Figure 6: Relationship between prior expectation (the average of historical ratings) and the average of the next rating. A smaller d implies a better fitting.

Due to the page limit, we only describe the experimental steps on HIALF here. Experiments on other models are similar. For each dataset, we first model it using HIALF. Then, we select the same groups of products as in Figure 1. In each selected group, for each rating, given its history $\mathcal{H}_{p,i}$, we use HIALF to predict the next rating $r^{HIALF}_{p,i}$. We also calculate its prior expectation $e_{p,i}$ based on real ratings, getting one pair $(e_{p,i}, r^{HIALF}_{p,i})$. We consider two types of $e_{p,i}$ here: (1) the average of historical ratings; (2) the one defined in Equation (7) that focuses more on recent ratings. Finally, for each type of $e_{p,i}$, we plot the relationship between prior expectation and the average of the next HIALF-generated rating. Figure 5 shows the results with prior expectation defined in Equation (7), while prior expectation in Figure 6 is the average of historical ratings. Here, we only plot the relationship for the group of products with average ratings in $[3.9, 4.1]$ in each dataset because this group contains more products. Similar patterns are also found in other groups of products.

Summary of results. Both Figure 5 and Figure 6 indicate that HIALF provides the best fit to previous observations in real ratings. Take Figure (5a) and Figure (6a) as examples. The black lines with ○ are the relationship between prior expectation and the average of the next rating in real ratings, and we can find our model HIALF (red line with □) fits the relationship of real ratings the best, as compared to LF (blue line with △) and HEARD (the purple line with ×). We also define a quantative metric to measure the difference between relationship in real ratings and in ratings generated by model A (where A can be HEARD/LF/HIALF) as: $d = \frac{\sum_{e \in E}(r_e - r^A_e)^2}{|E|}$, where E contains all distinct prior expectation e, r_e is the average of real ratings under e, and r^A_e is the average of model A-generated ratings under e. HIALF also reveals the smallest d, implying the closest fitting to empirical observations in real ratings. The latent factor model (LF) reveals relationships that are approximately parallel to the x axis, since LF does not consider the factors of distortions from historical ratings. HEARD is too optimistic since it always tends to predict high ratings when prior expectations are larger than 3.

5 APPLICATIONS

In this section, we apply HIALF to improve recommendations and to help users to make wiser consuming decisions.

5.1 Debiased Recommender System

Using HIALF, one can easily obtain users' and products' intrinsic features ($b_p, b_u, \boldsymbol{x_u}, \boldsymbol{y_p}$) without any contamination from historical ratings. Thus, base on these intrinsic features, for a product p that user u has not consumed, we can generate a recommendation score:

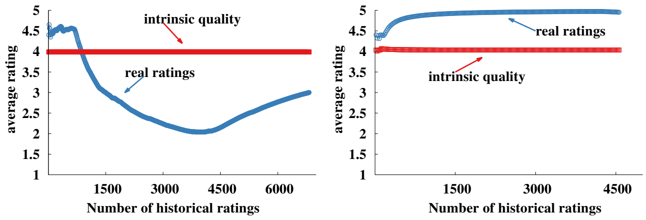
$$rec(p, u) = g + b_p + b_u + \boldsymbol{x_u}^T \boldsymbol{y_p} \quad (8)$$

here $g, b_p, \boldsymbol{x_u}, \boldsymbol{y_p}, b_u$ are learned parameters in HIALF. Products with high recommendation scores are those potential products that user u may like, and therefore we recommend these products to user u. We call a recommender system using the above methodology as *debiased recsys*.

We compare *debiased recsys* with the standard latent factor model (LF) since HIALF is built on top of the latent factor model. Note that HIALF is orthogonal to other techniques to improve recommendations, such as modeling evolution of users' expertise [17], modeling

Table 4: RMSE on five datasets

category	LF	debiased recsys
Amazon-movie	1.0639	**1.0465**
Amazon-books	0.9125	**0.8922**
Amazon-electronics	1.2273	**1.2083**
Amazon-clothes	1.1239	**1.1034**
Tripadvisor	1.1919	**1.1776**



(a) sample product 1 (b) sample product 2

Figure 7: Two products with similar intrinsic quality have different rating growth histories, leading to significantly distinct ratings.

temporal dynamics [12], etc. For the future work, we can combine HIALF with the above techniques for further improvements.

We take the set of ratings without historical ratings as the ground truth. We train HIALF with the rest of ratings using the same hyperparameters ($\lambda_\beta, \lambda_f, etc.$) as in Section 4.2. We report the Root Mean Square Error (RMSE) on the ground truth in Table 4. RMSE is widely used to evaluate the quality of recommendations [8, 11, 12].

As in Table 4, *debiased recsys* consistently reveals smaller RMSE than LF, implying that it can provide better recommendations.

5.2 Exposing The Intrinsic Product Quality

The intrinsic quality of a product is around the aggregated collective ratings given by a large group of users who were not exposed to historical ratings [25]. With HIALF, we can also easily get the intrinsic quality of product p, which we denote as q_p^*, by factoring out the distortions from historical ratings.

$$q_p^* = \sum_i (g + b_p + \mathbf{x}_{\tilde{u}(p,i)}{}^T \mathbf{y}_p) \tag{9}$$

Here $\tilde{u}(p, i)$ is the user who gave the i-th rating of product p.

We use the case study in Figure 7 to illustrate the significance of revealing the intrinsic qualities of products. Figure 7 shows the dynamics of the average rating of two selected products in *Amazon-movie*. These two products have similar intrinsic quality (around 4) and similar initial ratings. Note that initial ratings suffer small historical distortions. However, after they experienced a sequence of ratings with different trends, the average rating of product 1 and product 2 are 3.2 and 4.9 respectively (differ at about 1.7). This shows the impact of historical ratings' distortions. With HIALF, one can perform debiasing operation and obtain the intrinsic quality so that users will not be misguided by historical ratings.

6 RELATED WORK

Biases in rating system. Users' ratings are often *biased*, due to a variety of causes, such as ratings from spammers [15] or water-armies [1], evolution of users' expertise [17], temporal dynamics [12], dimensional biases [9], biases across categories [10], biases due to algorithms [24] etc. In this paper, we focus on a different kind of bias caused by influence from historical ratings.

Experiments on historical ratings' influence. Recent studies [3, 19, 22, 30] found that the disclosed historical ratings would distort subsequent ratings. Experiments [19, 30] revealed that small positive manipulations would encourage more positive future ratings, creating accumulative herding that boosts the final average ratings. Even for products with the same quality, users tend to rate higher when they are displayed with higher historical ratings [3, 22]. Our work is motivated by the above findings, however, our goal is to model rather than to test the influence from historical ratings.

Modeling historical ratings' influence. Previous works [2, 13] have attempted to mitigate the *micro-level* influence from historical ratings. However, their models were developed for specially designed rating systems, and one needs to know users' ratings given when users cannot see historical ratings, which is usually latent in reality. Wang et al. [28] then developed a more practical model (HEARD) to characterize the macro-level influence from historical ratings on *Amazon*, i.e., how historical ratings of a product will affect its general rating distribution after 100 ratings. The goal is different from our work since we aim to capture the microscopic influence, i.e., how historical ratings will affect its next single rating.

Social network-based influence. Several works [5, 6] also modeled and debiased the influence in social network, i.e., *peer effects*. Peer effects are interactive and more credible, i.e., users and their friends will influence each other and users often trust each other. The historical ratings are usually generated by strangers, and only previous ratings can influence the subsequent ratings. The difference between these two types of influence makes our work differ from this line of works.

7 CONCLUSION AND FUTURE WORK

In this paper, using 42 million ratings from *Tripadvisor* and *Amazon*, we first reveal and explain the assimilation and contrast effects in users' given ratings caused by historical ratings. Then we propose HIALF, the first model for real rating systems to characterize the *micro-level* influence from historical ratings in each single rating. We demonstrate the effectiveness of HIALF in predicting subsequent ratings, capturing dynamics in real ratings, and providing better recommendations, and further revealing products' intrinsic qualities for subsequent wiser decisions on purchasing products.

There are several directions for future work. First, besides ratings, review texts also contain a lot of information. The recent work [16] has combined reviews and ratings for better recommendations. Thus one can further improve the HIALF model by incorporating useful information embedded in the review texts. Also, HIALF is orthogonal to other factors in ratings, such as evolution of user's expertise [17], temporal dynamics [12], etc. Considering these factors may contribute to a better model, we plan to do this in our future work.

ACKNOWLEDGMENTS

The authors would like to thank anonymous reviewers for their valuable comments. This work is supported in part by the GRF 14208816 and Huawei Research Grant.

REFERENCES

[1] 2017. Internet Water Army. https://en.wikipedia.org/wiki/Internet_Water_Army. (2017). (accessed March 2017).

[2] Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. 2014. De-biasing user preference ratings in recommender systems. In *IntRS Workshop@RecSys 2014*. 2–9.

[3] Gediminas Adomavicius, Jesse Bockstedt, Shawn P. Curley, and Jingjing Zhang. 2016. Understanding Effects of Personalized vs. Aggregate Ratings on User Preferences. In *IntRS Workshop@RecSys 2016*. 14–21.

[4] Rolph E Anderson. 1973. Consumer dissatisfaction: The effect of disconfirmed expectancy on perceived product performance. *Journal of marketing research* (1973), 38–44.

[5] Abhimanyu Das, Sreenivas Gollapudi, Rina Panigrahy, and Mahyar Salek. 2013. Debiasing social wisdom. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 500–508.

[6] Anirban Dasgupta, Ravi Kumar, and D Sivakumar. 2012. Social sampling. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 235–243.

[7] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 293–296.

[8] Rana Forsati, Iman Barjasteh, Farzan Masrour, Abdol-Hossein Esfahanian, and Hayder Radha. 2015. Pushtrust: An efficient recommendation algorithm by leveraging trust and distrust relations. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 51–58.

[9] Yong Ge and Jingjing Li. 2015. Measure and Mitigate the Dimensional Bias in Online Reviews and Ratings. (2015).

[10] Fangjian Guo and David B Dunson. 2015. Uncovering Systematic Bias in Ratings across Categories: a Bayesian Approach. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 317–320.

[11] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 426–434.

[12] Yehuda Koren. 2009. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 447–456.

[13] Sanjay Krishnan, Jay Patel, Michael Franklin, and Ken Goldberg. 2014. Social influence bias in recommender systems: a methodology for learning, analyzing, and mitigating bias in ratings. In *Proceedings of the 8th ACM Conference on Recommender systems*. 137–144.

[14] Bibb Latané. 1981. The psychology of social impact. *American psychologist* 36, 4 (1981), 343.

[15] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. 2010. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 939–948.

[16] Guang Ling, Michael R Lyu, and Irwin King. 2014. Ratings meet reviews, a combined approach to recommend. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 105–112.

[17] Julian McAuley and Jure Leskovec. 2013. From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews. In *WWW'13: Proceedings of the 22ed International World Wide Web Conference*.

[18] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.

[19] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.

[20] Richard L Oliver. 2014. *Satisfaction: A behavioral perspective on the consumer*. Routledge.

[21] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. *Introduction to recommender systems handbook*. Springer.

[22] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311, 5762 (2006), 854–856.

[23] Franklin E Satterthwaite. 1946. An approximate distribution of estimates of variance components. *Biometrics bulletin* 2, 6 (1946), 110–114.

[24] Patrick Shafto and Olfa Nasraoui. 2016. Human-Recommender Systems: From Benchmark Data to Benchmark Cognitive Models. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 127–130.

[25] James Surowiecki, Mark P Silverman, et al. 2007. The wisdom of crowds. *American Journal of Physics* 75, 2 (2007), 190–192.

[26] Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 618–626.

[27] Jian Wang and Yi Zhang. 2013. Opportunity Models for E-commerce Recommendation: Right Product, Right Time. In *SIGIR*.

[28] Ting Wang, Dashun Wang, and Fei Wang. 2014. Quantifying herding effects in crowd wisdom. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1087–1096.

[29] Larry Wasserman. 2013. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

[30] Tim Weninger, Thomas James Johnston, and Maria Glenski. 2015. Random voting effects in social-digital spaces: A case study of reddit post submissions. In *Proceedings of the 26th ACM conference on hypertext & social media*. ACM, 293–297.

[31] Ming Yan, Jitao Sang, and Changsheng Xu. 2015. Unified YouTube Video Recommendation via Cross-network Collaboration. In *ICMR*.

[32] Haiyi Zhu and Bernardo A Huberman. 2014. To switch or not to switch: understanding social influence in online choices. *American Behavioral Scientist* 58, 10 (2014), 1329–1344.