Understanding Assimilation-contrast Effects in Online Rating Systems: Modelling, Debiasing, and Applications

XIAOYING ZHANG, The Chinese University of Hong Kong
HONG XIE, Chongqing University
JUNZHOU ZHAO and JOHN C. S. LUI, The Chinese University of Hong Kong

"Unbiasedness," which is an important property to ensure that users' ratings indeed reflect their true evaluations of products, is vital both in shaping consumer purchase decisions and providing reliable recommendations in online rating systems. Recent experimental studies showed that distortions from historical ratings would ruin the unbiasedness of subsequent ratings. How to "discover" historical distortions in each single rating (or at the micro-level), and perform the "debiasing operations" are our main objective. Using 42M real customer ratings, we first show that users either "assimilate" or "contrast" to historical ratings under different scenarios, which can be further explained by a well-known psychological argument: the "Assimilate-Contrast" theory. 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 allows us to study the influence patterns of historical ratings from a modelling perspective, which perfectly matches the assimilation and contrast effects observed in experiments. Moreover, HIALF achieves significant improvements in predicting subsequent ratings and characterizing relationships in ratings. It also contributes to better recommendations, wiser consumer purchase decisions, and deeper understanding of historical distortions in both honest rating and misbehaving rating settings.

CCS Concepts: • **Information systems** → **Data mining**; • **Social and professional topics** → **User characteristics**;

Additional Key Words and Phrases: Modelling and debiasing historical ratings' influence, recommender systems

ACM Reference format:
Xiaoying Zhang, Hong Xie, Junzhou Zhao, and John C. S. Lui. 2019. Understanding Assimilation-contrast Effects in Online Rating Systems: Modelling, Debiasing, and Applications. *ACM Trans. Inf. Syst.* 38, 1, Article 2 (October 2019), 25 pages.
https://doi.org/10.1145/3362651

1 INTRODUCTION

Contemporary web services provide many important applications ranging from e-commerce websites [4, 6, 21, 31] to online video/news platforms [9, 36]. One of the most important modules of

This work is partially supported by GRF 14208816.
Authors' addresses: X. Zhang, J. Zhao, and J. C. S. Lui, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR; emails: xyzhang@cse.cuhk.edu.hk, junzhouzhao@gmail.com, cslui@cse.cuhk.edu.hk; H. Xie (corresponding author), Chongqing University, No.174 Shazhengjie, Shapingba, Chongqing; email: xiehong2018@cqu.edu.cn.

1046-8188/2019/10-ART2 $15.00
https://doi.org/10.1145/3362651

these web services is the online rating system. 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. 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, because they rely on the unbiasedness to make reliable recommendations.

Recent experimental studies [3, 22, 26, 34] showed that the disclosed historical ratings would ruin the *unbiasedness* of subsequent ratings, making them inaccurate to convey users' true (or intrinsic) evaluations of 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 [22, 26, 34]. At the micro level (or at the granularity of each single rating), the distortion in the rating provides an adulterated view of a user's preference for the product, weakening recommender systems' ability to provide high-quality recommendations [3]. As in Reference [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 gives a high rating to a product under high historical ratings, the high rating may not suggest the user's high preference to the product anymore, since this high rating may be influenced by the high ratings in the past.

Recently, Wang et al. [32] 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, 16] tried to mitigate the micro-level historical ratings' influence with an assumption that we know *users' intrinsic ratings*, the ratings given when users could not 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 may influence subsequent rating behavior.

The goal of this work is to develop a novel 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 the next rating.

In particular, we first analyze a dataset of 42M 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 [5] in psychology. Then, we found that previous works (References [2, 16, 32]) were unable to explain our empirical results. Thus, we propose the *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 [32] and 12% as compared with the standard latent factor model [25].

To show the utility of HIALF, we apply the model to two applications. 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 misled by distorted historical ratings.

Last, we also conduct several analyses with HIALF. We first analyze the strength of historical distortions on different datasets. Such analysis provides further justifications to our previous experimental results and gives guidance on applying HIALF. Moreover, we extend our model to understand how misbehaving/fake ratings will affect subsequent rating behavior. We conduct extensive experiments and reveal important insights on when a small number of misbehaving ratings can distort subsequent product ratings significantly and lead the average rating of a product to diverge. These observations can be applied to design effective misbehaving rating detection/defending algorithms.

Contributions. Overall, we make the following contributions:

- **Experimental observations.** We first reveal the assimilation and contrast effects in users' rating behavior caused by historical ratings. We also provide an explanation for our observations by a well-known psychological theory.
- **Mathematical modelling.** We develop the first model (HIALF) for real rating systems to characterize and mitigate historical distortions in each single rating microscopically.
- **Accuracy of our model.** The discovered influence patterns of historical ratings via HIALF perfectly match the assimilation and contrast effects in observations. Moreover, HIALF achieves significant improvements in predicting subsequent ratings, and accurately fits the relationships revealed in empirical measurements on real ratings.
- **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.
- **Analyses.** Important insights are revealed through the analysis of historical distortions on different datasets and the analysis of misbehaving ratings' impacts.

This article is organized as follows: Section 2 presents a variety of previous experimental studies, which justified that historical ratings influence subsequent ratings. Section 3 presents our analysis of rating datasets from Amazon and TripAdvisor. We observe the pattern of historical influence and give explanations of the pattern, i.e., how historical ratings influence the next rating. Section 4 presents our HIALF model to capture the historical influence. Section 5 presents the experiments to evaluate our model. Section 6 introduces two applications of HIALF model, and Section 7 presents the analysis of historical distortions on different datasets and the analysis of misbehaving ratings' impacts. Section 8 discusses related works, and Section 9 concludes our article.

2 BACKGROUND: EVIDENCE OF HISTORICAL INFLUENCE

In this section, we briefly introduce several previous experimental studies, which showed that the disclosed historical ratings influence subsequent ratings. These studies build the foundation of our work, i.e., they show that the historical ratings truly influence subsequent ratings. However, they did not show how the historical ratings will influence subsequent ratings, which is the main goal of our work.

Recently, a variety of experimental studies were conducted on different platforms to study the side effects caused by disclosing historical ratings. Typically, in these experimental studies, users' rating behaviors are randomly measured in two different scenarios: *showing historical ratings* or *hiding historical ratings*. By comparing users' behavior patterns under these two scenarios, they concluded that historical ratings influence subsequent ratings. Salganik et al. [26] implemented

a music lab, where users download and rate songs with or without information about how good the songs are. They demonstrated that increasing historical influence could result in different outcomes for songs with similar quality. Muchnik et al. [22] and Weninger et al. [34] conducted similar experiments on news platforms (e.g., Reddits), and found that small manipulations in historical ratings will create significant changes in downstream ratings, resulting in significantly different final outcomes. With a joking rating dataset, Adomavicius et al. [3] demonstrated similar results.

All the above experiments only found that the disclosed historical ratings influence subsequent user ratings, making them inaccurate to convey users' true evaluations. However, they did not show the specific pattern of how historical ratings influence subsequent ratings, and this is what we will do in the next section.

3 HOW HISTORICAL RATINGS AFFECT THE NEXT SINGLE RATING

To understand how historical ratings may affect the next rating, we first conduct empirical measurements on real-world datasets to study how historical ratings may impact the next single rating. In this section, we first describe these rating datasets, then we discuss how to measure the impact of historical ratings on the next rating. Finally, we propose an explanation of our empirical observations, and verify that the existing works [2, 16, 32] cannot explain our observations, which motivates us to design a new model for real rating systems to describe the micro-level historical ratings' influence.

3.1 Rating Datasets

We first introduce two large-scale public available rating datasets from Amazon[1] and TripAdvisor,[2] respectively. Table 1 summarizes the basic statistics of our datasets.

(1) Dataset from Amazon: Amazon is a popular e-commerce website that allows users to review and rate products they recently consumed, e.g., books, clothes, and so on. In the Amazon dataset [21], 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.

(2) Dataset from TripAdvisor: TripAdvisor is a popular travel website that provides reviews and ratings of travel-related contents, e.g., hotels, restaurants, and so on. We use the entire ratings on it from April 2001 to September 2012 [30].

3.2 Empirical Measurements and Observations

To gain a better understanding, we conduct empirical measurements on the above datasets to study how historical ratings affect the next rating. Let $r_{p,i}$ denote the ith rating of product p, and let

$$\mathcal{H}_{p,i} \triangleq (r_{p,1}, \ldots, r_{p,i-1}) \tag{1}$$

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 $r_{p,i}$.

Our goal is to measure how historical ratings $\mathcal{H}_{p,i}$ affect the 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) historical ratings to which the user was exposed.

Note that the first factor is "*latent*" and around the average of ratings given by a large population who were not exposed to historical ratings [29]. To process our dataset, we group products with

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

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

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 leave out a statistically insignificant group containing fewer than 100 products, and on average, each dataset has 10 groups. Then, in each product group, for each rating $r_{p,i}$, we calculate its *prior expectation* formed on historical ratings $\mathcal{H}_{p,i}$:

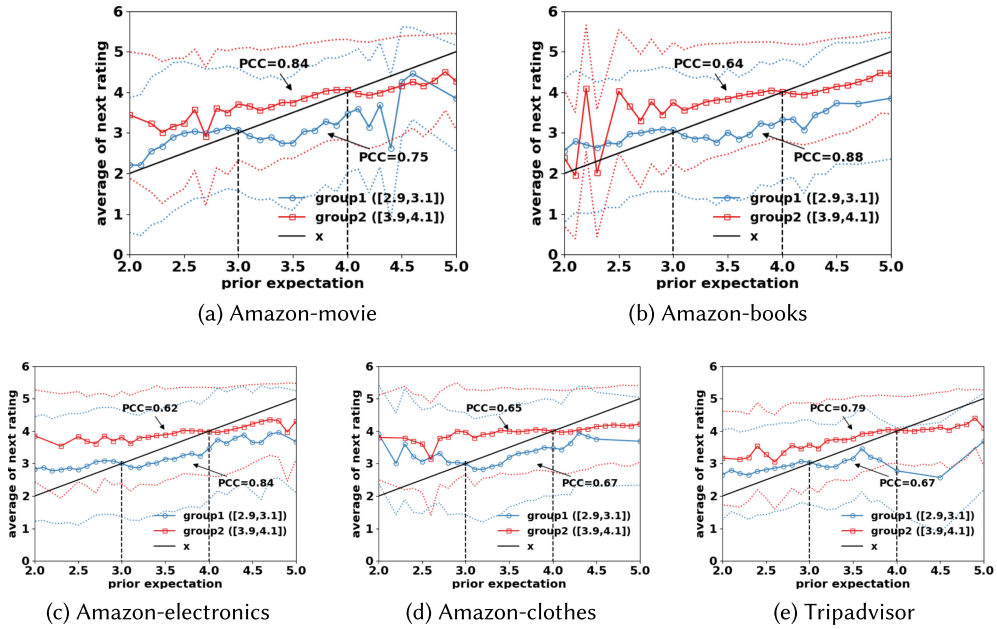
$$e_{p,i} = \frac{1}{i-1} \sum_{k=1}^{i-1} r_{p,k}, \tag{2}$$

resulting in a pair $(e_{p,i}, r_{p,i})$. To figure out the relationship between the prior expectation and the next rating, we follow the idea of binned scatterplots. More specifically, we round the prior expectation $e_{p,i}$ to one decimal place, i.e., each bin has a width of 0.1. Let the set E denote all distinct prior expectations after the rounding. And for any $e \in E$, a set of pairs $\{(e, \dot{r}_1), \ldots, (e, \dot{r}_{n_e})\}$ have the same prior expectation e, but the ratings $\{\dot{r}_1, \ldots, \dot{r}_{n_e}\}$ are given by different users. Here n_e denotes the number of ratings under prior expectation e. We aggregate ratings $\{\dot{r}_1, \ldots, \dot{r}_{n_e}\}$ to get $\bar{r}_e = \frac{1}{n_e} \sum_{i=1}^{n_e} \dot{r}_i$. Thus, the resulting list $\{(e, \bar{r}_e) | e \in E\}$ describes how prior expectation e affects the next rating $\bar{r}_e$, on average, in this product group. Finally, we plot the relationship between prior expectation e and the average of the next rating $\bar{r}_e$ for each selected product group in Figure 1. Relationships in other groups are similar with the two selected groups. In Figure 1, the solid line with mark □ (○) depicts the relationship between prior expectation e and the average of the next rating $\bar{r}_e$ in product group $[3.9, 4.1]$ ($[2.9, 3.1]$). Each solid line is accompanied with two dotted lines in the same color, which represent the 95% confidence interval for the corresponding product group.

The wide confidence interval as well as variations in plots are caused by users' personal preference. However, users' personal preference will not affect the plotted relationships between e and $\bar{r}_e$, because for each prior expectation e, we aggregate users' ratings under the same prior expectation to obtain the average of the next rating $\bar{r}_e$. When ratings under each prior expectation are given by many users, the aggregated distortions of those users' personal preference are insignificant. To ensure this, we leave out prior expectations followed by less than 100 ratings due to their low statistical reliability. As shown in Figure 1, the confidence interval in two product groups is roughly stable over different prior expectations, which means the bias of users' personal preference is well controlled.

Examining the results via Figure 1, we made the following observations:

- Products' historical ratings do affect the next rating. For example, in Figure 1(a), 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 0.75 and 0.84, respectively. In general, the Pearson correlation coefficients (PCC) are in the range $[0.62, 0.88]$, which reflects a positive correlation between prior expectation and the next rating.



(a) Amazon-movie (b) Amazon-books

(c) Amazon-electronics (d) Amazon-clothes (e) Tripadvisor

Fig. 1. (a)–(e): Relationship between prior expectation e and the average of the next rating $\bar{r}_e$ in different datasets. 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 the 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).

- 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 the next rating, i.e., the user will give a 4-star rating as long as his/her prior expectation is 4. Take the group 2 in Figure 1(a) 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. We would like to emphasize that this phenomenon is "*consistent*" among all groups of products in our datasets, and it is pretty interesting to find a model that can accurately explain this phenomenon.

More remarks on the observations:

- We do not use the first observation to argue the existence of historical influence. Actually, as mentioned in Section 2, the existence of historical influence has already been certified by previous experimental studies. And our first observation coincides with the results of previous experimental studies.
- Because we choose the product group by products' average rating, one may suspect the observations are just a mathematical property of the experimental design. However, if the observations are just a mathematical property of average ratings, there can be many possible patterns; for example, always receiving a rating equal to the group's approximately true quality under different expectations, or receiving a higher rating when prior expectation is higher than the group's approximately true quality, while a lower rating when prior

Table 2. Information about Linear Regression

category	slope		p-value	
	group 1	group 2	group 1	group 2
Amazon-books	0.39	0.38	5.2e-12	6.5e-05
Amazon-clothes	0.34	0.14	7.5e-05	9.4e-05
Amazon-electronics	0.32	0.14	3.3e-09	2.9e-04
Amazon-movies	0.46	0.33	3.5e-06	2.8e-09
TripAdvisor	0.20	0.29	1.1e-04	5.6e-08

expectation is lower than the group's approximately true quality. Why is there one specific pattern in our observations that is *consistent* among all groups of products in datasets?

- Note that if historical ratings, i.e., prior expectations, have no effects on subsequent ratings, the relationship between the prior expectation and the average of the next rating should be parallel to x axis with slope = 0. To make our observations clearer, we do linear regression for each plotted relationship and record the slope of fitted lines in Table 2. Moreover, we do a hypothesis test whose null hypothesis is that the slope is zero. Table 2 also shows the two-sided p-value for the hypothesis test. Both non-zero slopes and the p-value smaller than 0.05 certify again that historical ratings truly have influence on subsequent ratings.

3.3 Proposed Explanation of Observations

Next, we aim to answer the following two fundamental questions: (1) *Why do historical ratings influence the next rating?* (2) *Why does the influence of historical ratings behave consistently like those in Figure 1?*

For the first question, one possible answer 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/her needs. At this stage, he/she forms his/her "prior expectation" for that product. Using the customer satisfaction theory [24], a 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, making the next single rating different.

For the second question, we refer to three well-known psychological theories [5] that 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 2(a) shows the sample representations of three theories. The product quality is 3 and is represented by the line parallel with x axis.

(1) **"Assimilate" theory:** The user's satisfaction of the product is always similar to his/her prior expectation (the orange line with △).
(2) **"Contrast" theory:** The customer will magnify the difference between his/her prior expectation for the product and the product quality; i.e., if his/her 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 ○).
(3) **"Assimilate-Contrast" theory:** If the disparity between his/her prior expectation and the product quality can be accepted by the user (in $[\theta, \sigma]$ in Figure 2(a)), then 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 □).

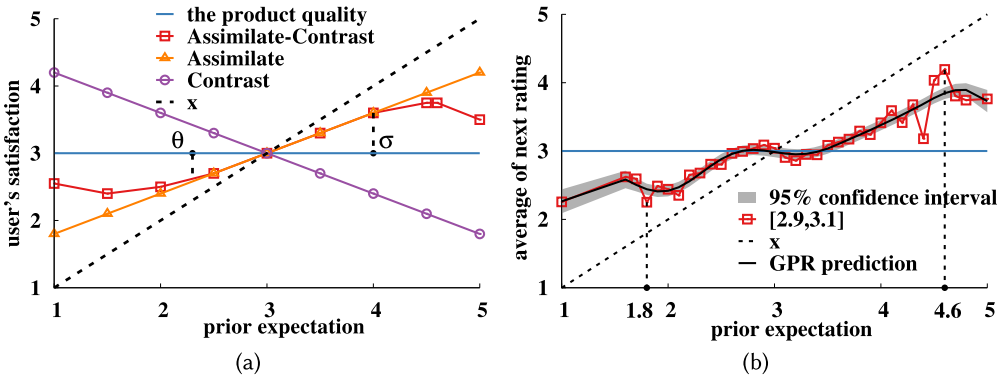


Fig. 2. (a) Sample representations of three theories. (b) Relationship between prior expectation and the average of the next rating in products with average ratings in $[2.9, 3.1]$ in all five datasets.

One important question we need to answer is which theory can explain our empirical observations best. To answer this question, we combine groups with the same average rating range in five datasets as a product group, and do analysis on each product group. We do two types of analysis: (1) Gaussian process regression; (2) statistical hypothesis test. And we show that the "Assimilate-Contrast" theory is more approximate to explain our empirical observations.

Gaussian process regression. We do a Gaussian process regression (GPR) on relationship of each product group. For example, Figure 2(b) illustrates the relationship for all products with average ratings in $[2.9, 3.1]$ as well as Gaussian process regression on it. The black line is the relationship predicted by GPR, and the shaded area is 95%-confidence interval of GPR prediction. We use this product group to illustrate, because its approximate true quality is in the middle of the rating scale of *Amazon* and *Tripadvisor*, i.e., $[1, 5]$. One can see that the relationship follows the *"Assimilate-Contrast"* theory, since:

- The prior expectation has a positive correlation with the next rating (contradicting with the *"Contrast"* theory).
- When the prior expectation is far away from its approximate true quality (i.e., in the range of $[1, 1.8)$ and $(4.6, 5]$), the average of the next rating as well as its GPR prediction diverge from the increasing trend when prior expectation is close to its approximate true quality, i.e., in $[1.8, 4.6]$. (contradicting with the *"Assimilate"* theory).

GRP regression on the other product groups suggest the same result. Note that the curve in Figure 2(b) is not very smooth, implying some variations. This is because the number of ratings under some prior expectations—for example, the ones far from product quality—are not that large. To handle this uncertainty, we leave out prior expectations followed by less than 100 ratings due to their low statistical reliability. We also try larger thresholds—for example, 150 or 200—so the resulted plotted curve is smoother but filters out more points.

Statistical hypothesis test. Next, we test whether the *"Assimilate-Contrast"* theory can explain our observations through a statistical hypothesis test. 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 $(\sigma, 5]$ as shown in Figure 2(a). 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 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.,

$$S_i = \left\{ \frac{r_e - \bar{r}_\theta}{e - \theta} | (e, r_e) \in C_i \right\}, \quad i = 1, 2. \tag{3}$$

For the above equation, $\bar{r}_\theta$ 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, then 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 $[a - \epsilon, a + \epsilon]$ are taken as the same prior expectation as a. In our experiments, we set $\epsilon = 0.1$. Then, we get $S_i, i = 1, 2, 3, 4$ and apply the student t-test. We use Welch-Satterthwaite approximation [27] 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 hypotheses 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"* model is a more appropriate theory to explain how historical ratings can affect customers in their subsequent ratings.

3.4 Limitations of Existing Works

Let us now show how existing works on modelling historical ratings' influence [2, 16, 32] fall short in explaining our previous observations.

Note that in our dataset, users' intrinsic ratings are latent, thus we are unable to build the model in References [2, 16]. Wang et al. [32] developed HEARD to model how historical ratings $\mathcal{H}_{p,i}$ influence the general rating distribution after next M ratings, denoted as $x_{p,i+M}$, 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, and we denote it as r^H_e.

Before checking the slope changes as in previous hypothesis tests, we first check whether the resulting list $\{(e, r^H_e)\}$ meets $r^H_e \leq e$, when $e \geq q$, as in Figure 1. Here q refers to the approximately

Table 3. 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

true quality of the product group. Let E_q^+ denote those prior expectations larger than q: $E_q^+ = \{e|e - q \geq 0\}$. We calculate the average deviation from e to r_e^H when $e \geq q$:

$$d^H = \frac{1}{|E_q^+|} \sum_{e \in E_q^+} (e - r_e^H).$$

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 $e \geq q$, which we denote as d^*. Note that d^* is always positive in the real rating datasets, because $r_e \leq e$, when $e \geq q$ in Figure 1. We present d^H and d^* in both groups on all five datasets in Table 3. From Table 3, 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_e^H 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 limitation of HEARD is due to the fact that HEARD mainly focuses on the macro-level historical ratings' influence in overall rating distribution, rather than the micro-level historical influence in each single subsequent rating.

4 HIALF MODEL

Let us now describe in detail the *Historical Influence Aware Latent Factor Model* (HIALF) that 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 to do a better prediction of the next rating; (2) to reveal the intrinsic qualities of products and users' intrinsic preference to do a better job in product recommendations.

4.1 Preliminary: Latent Factor Model

One can first consider using the classical latent factor (LF) model [25] to predict the rating $r_{u,p}$ for user u and product p as

$$r_{u,p} = g + b_u + b_p + \boldsymbol{x_u}^T \boldsymbol{y_p}. \tag{4}$$

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.

It is important 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, i.e., it extends LF model to incorporate the influence from historical ratings.

4.2 Historical Influence Aware Latent Factor (HIALF) Model

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 ith 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

$$q_{u,p} = g + b_p + \boldsymbol{x_u}^T \boldsymbol{y_p}. \tag{5}$$

Based on the 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 [17], 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

$$h_{p,i} = f(|\mathcal{H}_{p,i}|)\boldsymbol{\beta}(e_{p,i} - q_{u,p}). \tag{6}$$

In summary, HIALF predicts $\hat{r}_{p,i,u}$ for the ith rating of product p given by user u as follows:

$$\begin{aligned} \hat{r}_{p,i,u} &= b_u + q_{u,p} + \alpha_u h_{p,i} \\ &= g + b_u + b_p + \boldsymbol{x_u}^T \boldsymbol{y_p} + \alpha_u f(|\mathcal{H}_{p,i}|)\boldsymbol{\beta}(e_{p,i} - q_{u,p}). \end{aligned} \tag{7}$$

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 $e_{p,i}$.

Modelling 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 a 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 [33] to model $\boldsymbol{\beta}(x)$.

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

$$g_k(x) = \frac{\sum_{i=1}^n w(x, x_i) \cdot y_i}{\sum_{i=1}^n w(x, x_i)}. \tag{8}$$

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:

$$\boldsymbol{\beta(x)} = \frac{\sum_{l=1}^n w(x, e_l) \cdot v_l}{\sum_{l=1}^n w(x, e_l)}. \tag{9}$$

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 datasets, 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.

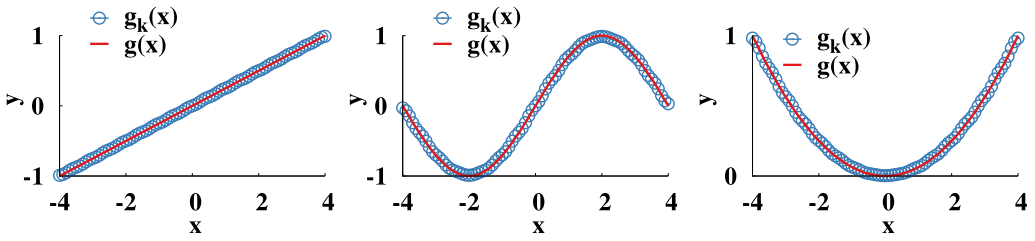


Fig. 3. Using kernel function to approximate three different functions with x in $[-4, 4]$: (a) $g(x) = \frac{x}{4}$; (b) $g(x) = \sin(\frac{\pi x}{4})$; (c) $g(x) = (\frac{x}{4})^2$. In kernel function, we use $\{x_1, \ldots, x_n\} = \{-4, -3.5, -3, \ldots, 3, 3.5, 4\}$ and $\kappa = 10$.

Modelling magnifying curve $f(x)$. Intuitively, the more historical ratings exist, the larger the magnifying effect will be. The previous psychological study [17] 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:

$$f(x) = \frac{a}{1 + exp(-b * x)} - \frac{a}{2}. \tag{10}$$

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.

Modelling 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:

$$e_{p,i} = \frac{\sum_{k=1}^{i-1} \xi(i-k) \cdot r_{p,k}}{\sum_{k=1}^{i-1} \xi(i-k)}. \tag{11}$$

Here, $\xi(d) = exp(-\gamma * d)$ denotes an exponential triggering kernel that models the decay of influence; $r_{p,k}$ is the kth 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 through experiments on a validation dataset.

4.3 Model Inference

Our goal is to solve the following optimization problem:

$$\min_{\Theta} \sum_{(p,i,u)\in\mathcal{K}} (r_{p,i,u} - \hat{r}_{p,i,u})^2 + \lambda_{rec}(b_u^2 + b_p^2 + ||\boldsymbol{x_u}||_2^2 + ||\boldsymbol{y_p}||_2^2)$$
$$+ \lambda_f(a^2 + b^2) + \lambda_\beta \left(\sum_l v_l^2 \right) + \lambda_\alpha(\alpha_u^2). \tag{12}$$

Here, $\Theta = \{g, \{b_u\}, \{b_p\}, \{\boldsymbol{x_u}\}, \{\boldsymbol{y_p}\}, \{\alpha_u\}, a, b, \{v_l\}\}$ and $r_{p,i,u}$ is the real rating. The $\hat{r}_{p,i,u}$ is the predicted rating by HIALF (Equation (7)), $\mathcal{K}$ contains all (p, i, u) pairs, and the (p, i, u) represents that the ith rating of product p in the dataset is given by user u. Of the parameters to learn, $g, a, b, \{v_l\}$ are global parameters, and $\{\boldsymbol{x_u}\}, \{\alpha_u\}, \{b_u\}$ are user-specific (i.e., each user has his/her own set of parameters), and $\{b_p\}, \{\boldsymbol{y_p}\}$ are product-specific (i.e., each product has its own set of parameters). To make it more clear, if we set $\{e_1, \ldots, e_n\} = \{-4, -3.5, \ldots, 3.5, 4\}$ and set the dimension of latent

Table 4. *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%

features as 5, we have $1 + 1 + 1 + 17 = 20$ global parameters, $5 + 1 + 1 = 7$ user-specific parameters for each user, and $1 + 5 = 6$ product-specific parameters for each product.

We use L2 loss function to make $\hat{r}_{p,i,u}$ as close as possible to the real rating $r_{p,i,u}$. Since there is no need for sparse output, we leverage L2 regularization due to its computational efficiency following previous works [20]. The $\lambda_{rec}, \lambda_f, \lambda_\beta, \lambda_\alpha$ are regularization hyperparameters to prevent overfitting. To learn the parameters, we use the stochastic gradient descent (SGD) algorithm, which has been widely used in previous works [10, 14, 15] because of its efficiency and accuracy.

5 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.

5.1 Predicting Subsequent Ratings

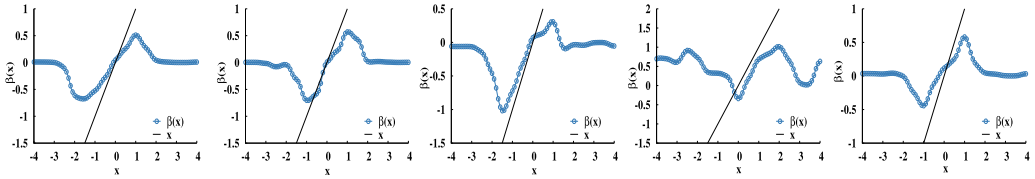
Experimental Setup. For the rating sequence of each product, we split it into the testing subsequence (the last 25 ratings) and the training subsequence (the rest of the ratings), and put the two subsequences into the training set and the testing set, respectively. Such construction avoids test-set biasing towards part of products—for example, products that have a large amount of ratings. We also remove the products with less than 50 training ratings, and combine users with fewer than 50 ratings as a big user. After the pre-processing, for example, the *Amazon-books* dataset contains 29,296 products, 6,450 users, and 8,193,695 ratings. We train the model on the training set, and validate the model on the testing set in terms of mean squared error (MSE), which is widely used in evaluating the accuracy of rating prediction [19, 20]. Smaller MSE suggests the higher accuracy of prediction. Specifically, the MSE on the test set $\mathcal{T}$ is defined as

$$MSE(\mathcal{T}) = \frac{\sum_{r_{p,i,u} \in \mathcal{T}} (r_{p,i,u} - \hat{r}_{p,i,u})^2}{|\mathcal{T}|}, \tag{13}$$

where $r_{p,i,u}$ is the real rating, and $\hat{r}_{p,i,u}$ is the predicted rating. To set hyperparameters in each model, we pick out each product's last 25 ratings in the training set and construct a validation set. We use the set of hyperparameters that reveal the best performance on the validation set. Following previous works [19, 20], we choose the dimension of latent features as 5.

We compare HIALF with several state-of-the-art models: latent factor (LF) model [25], HEARD [32], 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.

Results. Table 4 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 the improvements over LF are bounded by the extent to which ratings in the dataset



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

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

are affected by historical ratings as shown in Section 7.1. Furthermore, HIALF is consistently more accurate than HIALF-AVG, because users focus more on recent ratings when shaping prior expectations.

5.2 Validating the Disconfirmation Bias Curve

We then seek to validate whether the disconfirmation bias curve $\boldsymbol{\beta}(x)$ meets with the *"Assimilate-Contrast"* theory, because this will dictate the accuracy of HIALF.

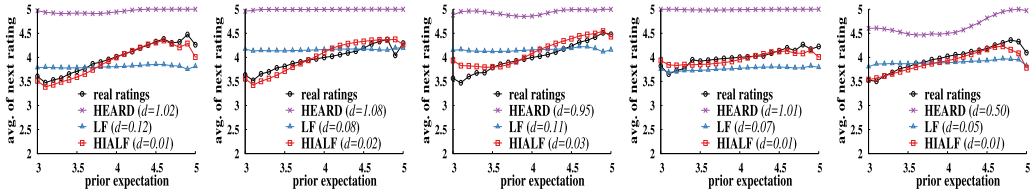
As in Figure 4, all learned $\boldsymbol{\beta}(x)$ match the *"Assimilate-Contrast"* theory well. $\boldsymbol{\beta}(x)$ on *Amazon-books*, *Amazon-movie*, *Amazon-electronics*, and *Tripadvisor* have similar formats with the sample representation of the *"Assimilate-Contrast"* theory in Figure 2(a). $\boldsymbol{\beta}(x)$ 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. This different $\boldsymbol{\beta}(x)$ suggests that users only follow prior expectation when prior expectation is above and not too far from their experienced quality when consuming clothing.

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 4(a). 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 Reference [38], if users find others' opinions highly contradict their own opinions, they may tend to insist on their own opinions.

5.3 Fitting Empirical Observations

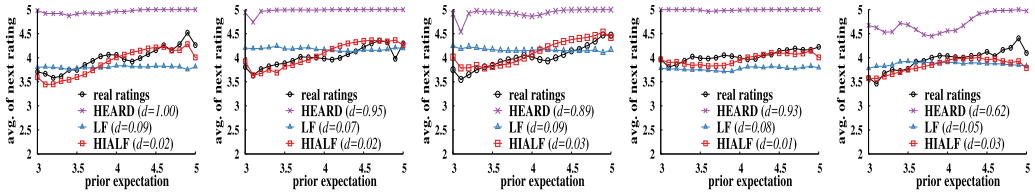
Next, we re-do the empirical measurements in Section 3 with the predicted ratings by HEARD, LF, HIALF, respectively. Note that an accurate model should reveal a similar relationship as in our previous observations in real ratings.

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 $r_{p,i}^{HIALF}$. We also calculate its prior expectation $e_{p,i}$ based on real ratings, getting one pair $(e_{p,i}, r_{p,i}^{HIALF})$. We consider two types of $e_{p,i}$ here: (1) the average of historical ratings; (2) the one defined in Equation (11) 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 (11), 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.



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

Fig. 5. Relationship between prior expectation (defined in Equation (11)) 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

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

Summary of results. Both Figure 5 and Figure 6 indicate that HIALF provides the best fit to previous observations in real ratings. Take Figure 5(a) and Figure 6(a) 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 quantitative 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:

$$d = \frac{\sum_{e \in E} (r_e - r_e^A)^2}{|E|}, \tag{14}$$

where E contains all distinct prior expectation e, r_e is the average of real ratings under e, and r_e^A 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.

5.4 Evaluation on Hyperparameters Testing

Our model is based on a set of hyperparameters, which are divided into two types:

- Regularization hyperparameters: $\lambda_{rec}, \lambda_f, \lambda_\beta, \lambda_\alpha$ in Equation (12), which are used mainly for preventing overfitting.
- Model-related hyperparameters: κ is used in modelling the disconfirmation bias curve $\boldsymbol{\beta}(x)$ (Equation (9)) and γ is used in modelling prior expectation (Equation (11)).

We next vary the above hyperparameters to study their effects. We use the same training set and test set as in Section 5.1, and take the hyperparameters used in Section 5.1 as default hyperparameters.

Table 5. *MSE* with Different λ_{rec}

λ_{rec}	0.5	0.1	0.05	0.01	0.005	0.001
Amazon-movie	1.1130	**1.1194**	1.1365	1.2015	1.1928	1.1804
Amazon-books	1.0284	1.0292	**1.0318**	1.0514	1.0583	1.0720
Amazon-electronics	1.8677	**1.8764**	1.8870	1.8986	1.9096	1.9130
Amazon-clothes	1.3718	1.3778	**1.3759**	1.3755	1.3767	1.3943
TripAdvisor	0.9920	1.0028	0.9829	**0.9405**	0.9428	0.9484

Table 6. *MSE* with Different λ_f

λ_f	0.01	0.001	0.0005	0.0001	0.00005	0.000001
Amazon-movie	1.1588	1.1229	1.1206	**1.1194**	1.1187	1.1185
Amazon-books	1.0765	1.0503	0.0418	1.0330	**1.0318**	1.0313
Amazon-electronics	1.9571	1.9413	1.8880	**1.8764**	1.8703	1.8717
Amazon-clothes	1.3867	1.3768	1.3764	**1.3759**	1.3759	1.3759
TripAdvisor	0.9512	0.9433	0.9430	**0.9405**	0.9402	0.9401

Effect of λ_{rec}. For each dataset, we vary λ_{rec} from 0.5 to 0.01, and fix other hyperparameters to be the same with experiments in Section 5.1. Table 5 shows the mean squared error (MSE) on five test datasets under different λ_{rec}. The bold part in each row is the MSE reported on each dataset in Section 5.1.

From Table 5, we can observe that the MSE on each test dataset varies with λ_{rec}. And when λ_{rec} decreases, the MSE on the test dataset becomes larger on all five datasets, i.e., prediction accuracy decreases. This is because smaller λ_{rec} leads to larger product quality $q_{u,p}$ and larger user biases $\{b_u\}$. And according to Equation (7), the factors modelling the historical influence will be smaller and thus lead to lower prediction accuracy. However, the MSE under different λ_{rec} are always smaller than the MSE of the standard latent factor model (LF) shown in Table 4. In other words, capturing the influence of historical ratings improves the prediction accuracy on subsequent ratings. And when λ_{rec} becomes larger on some dataset, e.g., *Amazon-books*, the MSE becomes smaller, while the MSE on *TripAdvisor* dataset becomes larger. This may be due to the difference of the dataset and the value of other hyperparameters.

Effect of λ_f. For each dataset, we vary λ_f from 0.01 to 0.00001, while fixing the other hyperparameters to be the same with that in Section 5.1. Table 6 shows the mean squared error (MSE) on each test dataset under different λ_f. The bold part in each row is the MSE reported on each dataset in Section 5.1.

From Table 6, we can observe that when λ_f is too large (i.e., $\lambda_f = 0.01$), the prediction accuracy decreases a lot. This is because too large λ_f makes $f(x)$ too small. Recall that $f(x)$ represents the magnifying effect of historical ratings with size x. And when λ_f is smaller than the default value, the prediction accuracy on all five datasets changes slightly. This may be due to the constraints from other hyperparameters.

Effect of λ_α. Similar as above, we vary λ_α from 0.1 to 0.0001, while fixing the other hyperparameters to be the same with that in Section 5.1. Table 7 shows the mean squared error (MSE) on five test datasets under different λ_α. The bold part in each row is the MSE reported on each dataset in Section 5.1.

From Table 7 one can observe that too large or too small λ_α can hurt the prediction accuracy. This is understandable, since too large λ_α leads to too small α_u, i.e., nearly ignoring the influence

Table 7. *MSE* with Different λ_α

λ_α	0.1	0.01	0.005	0.001	0.0005	0.0001
Amazon-movie	1.1198	**1.1194**	1.1228	1.1230	1.1228	1.1245
Amazon-books	1.0336	**1.0318**	1.0324	1.0352	1.0367	1.0370
Amazon-electronics	1.8863	1.8840	1.8806	**1.8764**	1.8766	1.8802
Amazon-clothes	1.3855	1.3745	1.3750	**1.3759**	1.3776	1.3832
TripAdvisor	0.9428	0.9421	0.9417	0.9408	**0.9405**	0.9410

Table 8. *MSE* with Different λ_β

λ_β	0.1	0.01	0.005	0.001	0.0005	0.0001
Amazon-movie	1.2095	1.1254	1.1210	1.1188	**1.1194**	1.1187
Amazon-books	1.0651	1.0510	1.0445	1.0339	**1.0318**	1.0310
Amazon-electronics	1.9628	1.9174	1.9020	1.8876	1.8827	**1.8764**
Amazon-clothes	1.4041	1.3833	1.3821	1.3793	1.3783	**1.3759**
TripAdvisor	0.9573	0.9473	0.9446	0.9411	**0.9405**	0.9410

Table 9. *MSE* with Different κ

κ	5	7	10	13	15	20
Amazon-movie	1.1199	1.1199	**1.1194**	1.1290	1.1269	1.1209
Amazon-books	1.0322	1.0324	**1.0318**	1.0317	1.0314	1.0325
Amazon-electronics	1.8710	1.8745	**1.8764**	1.8715	1.8723	1.8672
Amazon-clothes	1.3746	1.3755	**1.3759**	1.3767	1.3728	1.3754
TripAdvisor	0.9447	0.9424	**0.9405**	0.9421	0.9430	0.9498

from historical ratings, while too small λ_α leads to too large α_u, i.e., making the influence from historical ratings dominant.

Effect of λ_β. We vary λ_β from 0.1 to 0.0001, while fixing the other hyperparameters to be the same with that in Section 5.1. Table 8 shows the mean squared error (MSE) on five test datasets under different λ_β. The bold part in each row is the MSE reported in Section 5.1.

From Table 8 one can observe that when λ_β increases, the factor of historical influence becomes smaller, thus making the prediction accuracy decrease.

Effect of κ. We study the effect of κ used in modelling the disconfirmation bias curve $\boldsymbol{\beta}(x)$ (Equation (9)). Recall that κ controls the smoothness of $\boldsymbol{\beta}(x)$, and smaller κ corresponds to smoother curve. We vary κ from 5 to 20, while fixing the other hyperparameters to be the same with that in Section 5.1. Table 9 shows the mean squared error (MSE) on five test datasets under different κ. The bold part in each row is the MSE reported in Section 5.1. One can observe that the results change slightly with κ, i.e., they are roughly robust against the change of κ.

Effect of γ. The hyperparameters γ is used in modelling prior expectation (Equation (11)). Recall that γ controls the extent to which users prefer recent ratings. A larger γ means that users focus more on recent ratings. We vary γ from 0.01 to 0.1, while fixing the other hyperparameters to be the same with that in Section 5.1. Table 10 shows the mean squared error (MSE) on five test datasets under different γ. The bold part in each row is the MSE reported in Section 5.1.

From Table 10, one can find that when γ becomes larger, the prediction accuracy drops a lot. This means that users do not only focus on recent ratings too much. However, γ cannot be too small,

Table 10. *MSE* with Different γ

γ	0.01	0.03	0.05	0.07	0.09	0.1
Amazon-movie	1.1438	**1.1194**	1.1199	1.2049	1.2422	1.2499
Amazon-books	1.0467	1.0310	**1.0318**	1.0386	1.0436	1.0452
Amazon-electronics	1.8753	1.8642	**1.8764**	1.8910	1.8954	1.8968
Amazon-clothes	1.3842	1.3774	**1.3759**	1.3926	1.4077	1.4120
TripAdvisor	0.9703	0.9457	0.9427	0.9415	**0.9405**	0.9413

otherwise the prediction accuracy will also decrease, since users still prefer the recent ratings to some extent.

6 APPLICATIONS

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

6.1 Debiased Recommender System

Distortions from historical ratings weaken the system's ability to provide high-quality recommendations, since we cannot distinguish whether the given high rating is out of users' preferences or high historical ratings. Using HIALF, one can obtain users' and products' intrinsic features $(b_p, b_u, \boldsymbol{x_u}, \boldsymbol{y_p})$ without contamination from historical ratings. Thus, based on these intrinsic features, for a product p that user u has not consumed, we can generate a recommendation score:

$$rec(p, u) = g + b_p + b_u + \boldsymbol{x_u}^T \boldsymbol{y_p}. \tag{15}$$

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 modelling evolution of users' expertise [20], modelling temporal dynamics [15], and so on. For 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 (e.g., the first rating of each product). We train HIALF with the rest of the ratings using the same hyperparameters $(\lambda_\beta, \lambda_f, etc.)$ as in Section 5.1. We use two typical types of metrics to evaluate the accuracy of recommendations.

Root Mean Square Error (RMSE). RMSE is widely used to evaluate the accuracy of recommendations [10, 14, 15]. It quantifies the error of estimating users' intrinsic rating toward products. Small RMSE implies high recommencation accuracy. We report the RMSE on the ground truth in Table 11. As in Table 11, *debiased recsys* consistently reveals smaller RMSE than LF, implying that it can provide more accurate recommendations.

Relative Cumulative Reciprocal Rank (RCRR). Recommendation can be treated as a personalized preference ranking problem, i.e., ranking products based on a user's preferences (usually the inferred or estimated preferences) and then recommending top-ranked products to that user. Another typical way to quantify the accuracy of recommendation is through rank-based measures. As there are a large number of products, and the test set only contains a small fraction of products for each user, i.e., products that the user rated without historical ratings, we therefore define a new

Table 11. 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**

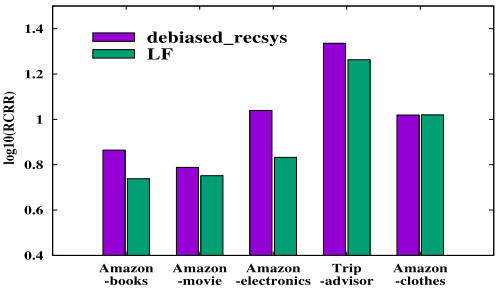


Fig. 7. Results in terms of RCRR in a log scale.

rank-based measure called "*relative cumulative reciprocal rank*" (RCRR) to quantify the recommendation accuracy. Let $\mathcal{D}_u$ be the set of items adopted by user u in the test set. Let $RCRR_u$ denote the relative cumulative reciprocal rank with respect to user u. Formally, we define $RCRR_u$ as

$$RCRR_u \triangleq \frac{1}{|\mathcal{D}_u|} \sum_{i \in \mathcal{D}_u} \frac{1}{(rank_{ui}/N)}, \tag{16}$$

where $N \in \mathbb{N}_+$ denotes the total number of products and $rank_{ui} \in \{1, \ldots, N\}$ denotes the rank of product i in the ranking list based on user u's preferences. For example, $rank_{ui} = 1$ means that product i is ranked as the top-1 for user u. $RCRR_u$ quantifies the average relative ranking of the products adopted by a user in the inferred ranking list (based on the inferred preference of this user) of products. For example, an $RCRR_u$ of 100 means that the items adopted by user u are in the top 1% of the inferred ranking list on average. Large $RCRR_u$ means that the adopted items are ranked higher on the list, i.e., implying a more accurate recommendation. The underlying intuition is that users tend to adopt products that match their preferences better. We define RCRR as the average of $RCRR_u$ over the whole user population. Figure 7 shows the *RCRR* for our debiased recsys and the LF algorithm in a log scale. One can observe that our debiased recommender system has a higher RCRR than LF, except for the *Clothes* dataset. This implies that debiased recommender system can make more accurate recommendations. For the *Clothes* dataset, our debiased recommender system has nearly the same RCRR as LF. This is because the user ratings in the *Clothes* dataset really suffer small distortions by historical ratings as suggested before.

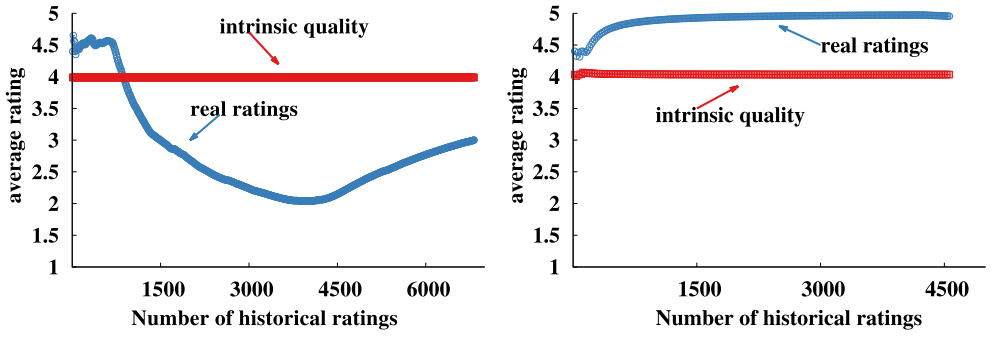
6.2 Exposing the Intrinsic Product Quality Using HIALF

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 [29]. With HIALF, we can also easily get the intrinsic quality of product p with N_p ratings, which we denote as q_p^*, by factoring out the distortions from historical ratings:

$$q_p^* = \frac{1}{N_p} \sum_{i=1}^{N_p} (g + b_p + \boldsymbol{x}_{\tilde{u}(p,i)}{}^T \boldsymbol{y}_p). \tag{17}$$

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

We use the case study in Figure 8 to illustrate the significance of revealing the intrinsic qualities of products. Figure 8 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 (differing by about 1.7). This shows the impact of historical ratings' distortions. With



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

Fig. 8. Two products with similar intrinsic quality have different rating growth histories, leading to significantly distinct ratings. The blue lines show the dynamics of the average of real ratings, and the red lines show the dynamics of the intrinsic quality extracted by HIALF.

HIALF, one can perform the debiasing operation and obtain the intrinsic quality so users will not be misguided by historical ratings.

7 ANALYSES OF RATING DISTORTIONS AND MISBEHAVIOR

In this section, we discuss more about historical distortions on different datasets. Moreover, we extend our model to study the impact of misbehaving ratings. We found that a small number of misbehaving ratings can distort subsequent product ratings significantly and make the average rating of a product diverge.

7.1 Rating Distortions on Different Datasets

We study the strength of historical distortions on different datasets, and how the length of historical ratings influences the strength of historical distortions on each dataset. In particular, we vary the length of historical ratings and calculate the corresponding average absolute historical distortions in ratings over all five datasets. Specifically, let

$$\mathcal{R}^p_{a,b} \triangleq \{r_{p,i} | an_p < i \leq bn_p\} \tag{18}$$

denote the rating subsequence between the an_pth rating and the bn_pth rating of product p, where $a, b \in [0, 1]$ and n_p is the total number of ratings of product p. For example, consider a product p has $n_p = 100$ ratings, then $\mathcal{R}^p_{0,0.2}$ denotes the set of the first 20 ratings of product p. To select the historical ratings of a dataset, we define

$$\mathcal{R}_{a,b} \triangleq \bigcup_{p \in \{\text{all products of a dataset}\}} \mathcal{R}^p_{a,b}. \tag{19}$$

Then, we calculate the average absolute historical distortions for all ratings in in $\mathcal{R}_{a,b}$ as:

$$\frac{\sum_{r_{p,i} \in \mathcal{R}_{a,b}} |\alpha_{\tilde{u}(p,i)} f(|\mathcal{H}_{p,i}|) \beta (e_{p,i} - q_{\tilde{u}(p,i),p})|}{|\mathcal{R}_{a,b}|}. \tag{20}$$

Note that ratings in $\mathcal{R}_{a,b}$ with larger a and b correspond to the scenario that a user is exposed to more historical ratings.

Figure 9 plots the average absolute historical distortions for $\mathcal{R}_{0,0.2}, \ldots, \mathcal{R}_{0.8,1}$. One can observe that with fixed a and b, items in *Amazon-movie* seem to suffer the largest historical distortions, followed by *Tripadvisor*, *Amazon-electronics,* and *Amazon-books*. The *Amazon-clothes* is the one

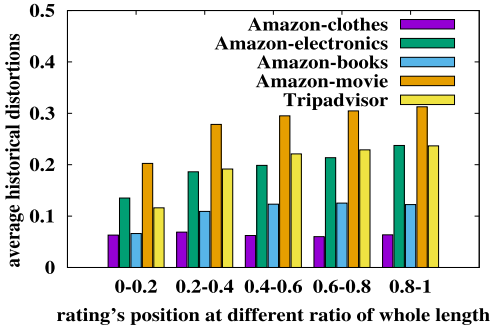


Fig. 9. Average absolute historical distortions.

with the smallest historical distortions. It is interesting to observe that in Table 4, HIALF has the most significant benefits on *Amazon-movie* over LF, followed by *Tripadvisor*, *Amazon-electronics*, and *Amazon-books*, and has the least benefits on *Amazon-clothes* over LF. In other words, the improvement of HIALF increases with the strength of historical distortions in the dataset.

Moreover, as we increase both a and b, the historical distortions in all datasets get larger. Namely, when a user is exposed to more historical ratings, his rating is more likely to be distorted more.

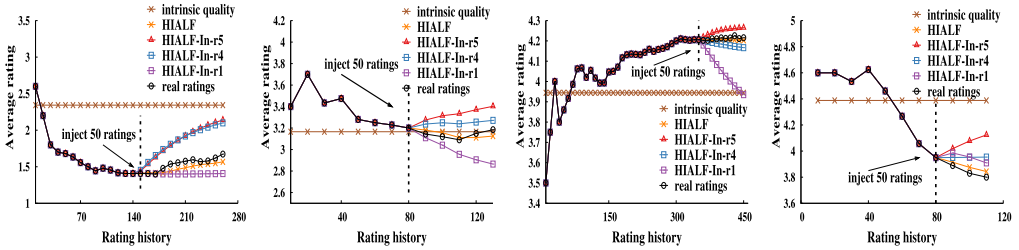
Lessons learned. Movie ratings are prone to historical distortions, while clothes ratings are not very prone to historical distortions. The strength of historical distortions in book ratings, electronics ratings, and hotel ratings lie between them. When a user is exposed to more historical ratings, his rating is more likely to be distorted more.

7.2 Misbehaving Ratings

Our results thus far consider honest ratings. However, in real-world online rating systems, there can be some misbehaving or fake ratings. Due to the openness of online rating systems, i.e., anyone can provide ratings to any product, some sellers or companies may hire users to provide high ratings intentionally to promote their products or even hire some users to provide low ratings to badmouth their competitors' products. In this section, we aim to extend our model to study the impact of such misbehaving ratings. In particular, we aim to answer: *How do misbehaving ratings influence the subsequent ratings of products?*

Misbehavior Model. Without loss of generality, we describe misbehaving ratings toward a single product. We use a tuple $(\bar{k}, L, \bar{N})$ to characterize misbehaving ratings, where $\bar{k}$ denotes the rating that misbehaving users provide, L denotes the length of misbehaving ratings (i.e., the number of misbehaving ratings), and $\bar{N}$ denotes the position of the first misbehaving rating. Note that we consider a consecutive sequence of misbehaving ratings. For example, under $(5, 3, 1)$, the first three ratings of a product would be $5, 5, 5$. Note that this misbehavior model is simple yet captures important factors of misbehaving ratings for our experimental studies.

Impact of misbehaving rating $\bar{k}$. In the following experimental studies, we choose four representative products denoted by A, B, C, D from our dataset to analyze, whose intrinsic quality range from low to high. We fix $L = 50$, i.e., we inject 50 misbehaving ratings to each product. We consider three cases of misbehavior with misbehaving rating $\bar{k} = 1, 4, 5$, respectively. For each case of misbehavior, the rating before the misbehavior is each product's historical ratings up to the misbehaving rating position, and we apply HIALF used in Section 6.2 to synthesize the subsequent ratings after each case of misbehavior.



(a) Product A (b) Product B (c) Product C (d) Product D

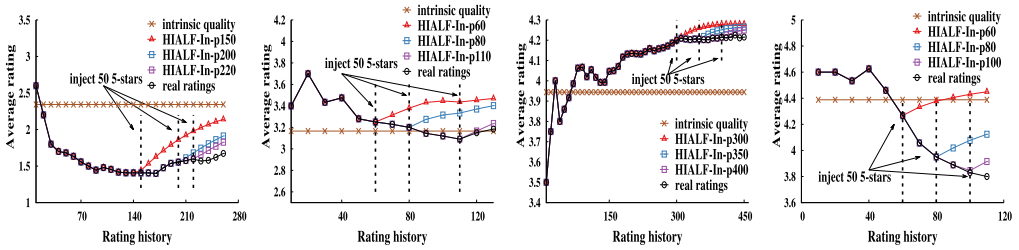
Fig. 10. Effects of the injected ratings.

The trends of a product's average rating after misbehavior is shown in Figure 10. In each subfigure of Figure 10, the orange line named HIALF represents the trend of average ratings synthesized by our HIALF model without artificial injections, and lines named HIALF-In-r5, HILF-In-r4, and HIALF-In-r1 represent the average rating trends synthesized by HIALF after inserting 50 5-star, 4-star, and 1-star ratings, respectively. The black line is the trend of average rating in real datasets, and the brown line parallel with x-axis is the intrinsic quality of the product.

From Figure 10(b) and Figure 10(c), one can observe that: (1) The higher the injected rating is, the larger the resulting average rating will be. For example, in both figures, the line named HIALF-In-r5 is above the line named HIALF-In-r4, which is above the line named HIALF-In-r1. (2) The line named HIALF-In-r1 is under the trend in real ratings and the trend generated by HIALF without injection. However, the first observation is not held in Figure 10(a) with HIALF-In-r5 and HIALF-In-r4 overlapped, while the second observation is not held in Figure 10(d) with HIALF-In-r1 above the two trends without artificial injection. Namely, the consequence of misbehaving ratings is not solely determined by the injected rating. Actually, it depends on both the injected rating and the intrinsic quality of the product. In Figure 10(a), the injected 5-star or 4-star ratings make users' prior expectation much larger than the intrinsic quality, and users prefer to insist their opinions and give ratings near the intrinsic quality. This is why the HIALF-In-r5 and HIALF-In-r4 overlap. Similarly, in Figure 10(d), the injected 1-star ratings make the users' prior expectation much lower than the product's intrinsic quality, and users also prefer to not assimilate to the prior expectations and give ratings near the intrinsic quality. Without misbehaving ratings, users will give a lower rating than the product's intrinsic quality, since they will assimilate to a prior expectation lower than the product's intrinsic quality. That is why trends in HIALF-In-r1 are higher than the two trends without misbehaving ratings. The intrinsic quality of products in Figure 10(b) and Figure 10(c) are in the middle of range $[1, 5]$, thus prior expectations after inserting misbehaving ratings are not too far from the product's intrinsic quality, and users will assimilate to prior expectations.

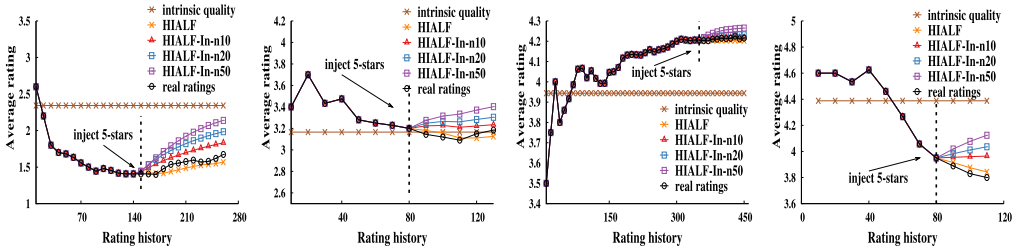
Impact of the injecting position $\bar{N}$. In this experiment, we fix $L = 50, \bar{k} = 5$, i.e., injecting 50 5-star ratings. We vary the injecting positions. The trends of products' average rating are shown in Figure 11. In each subfigure of Figure 11, the line named HIALF-In-p$\bar{N}$ denotes the trend of average rating after injecting the ratings just before the $\bar{N}$th rating. We can observe that the smaller $\bar{N}$ is, the larger the average rating will be in all subfigures. Namely, injecting ratings earlier results in larger impacts on subsequent ratings.

Impact of number of injected ratings L. In this experiment, we fix $\bar{k} = 5$, i.e., injecting 5-star ratings. We vary the number of ratings injected. We then apply the same steps as in above experiments to get the average rating trends under different scenarios and plot them in Figure 12. In each subfigure, the line named HIALF-In-nL represents the trend of average rating after



(a) Product A (b) Product B (c) Product C (d) Product D

Fig. 11. Effects of injecting position.



(a) Product A (b) Product B (c) Product C (d) Product D

Fig. 12. Effects of number of injected ratings.

injecting L ratings. From all subfigures, we can find that more injected ratings can distort subsequent ratings more.

Learned Lesson. Misbehaving ratings can distort the subsequent ratings significantly, and these distortions may lead the average rating of a product to diverge. This distortion is more significant when misbehaving ratings are injected earlier or a larger number of misbehaving ratings are injected.

8 RELATED WORK

Biases in rating system. Users' ratings are often *biased*, due to a variety of causes, such as ratings from spammers [18] or water-armies [1], evolution of users' expertise [20], temporal dynamics [13, 15, 35], dimensional biases [11], biases across categories [12], biases due to algorithms [28], and so on. In this article, we focus on a different kind of bias caused by influence from historical ratings.

Experiments on historical ratings' influence. Recent studies [3, 22, 26, 34] found that the disclosed historical ratings would distort subsequent ratings. Experiments [22, 34] 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, 26]. Our work is motivated by the above findings, however, our goal is to model rather than to test whether the influence from historical ratings exists.

Modelling historical ratings' influence. Previous works [2, 16] 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. [32] 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 the next single rating.

Social network–based influence. Several works [7, 8, 23, 37] also modeled and debiased the influence in social networks, 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.

9 CONCLUSION AND FUTURE WORK

In this article, using 42M 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, providing better recommendations, and further revealing products' intrinsic qualities for subsequent wiser decisions on purchasing products. Moreover, HIALF can also help us to gain a deeper understanding of historical distortions in normal ratings and effects of misbehaving ratings. Our model and observations not only can be applied to improve various rating-based applications such as recommendation, but also can be applied to design effective misbehaving rating detection/defending algorithms.

There are several directions for future work. First, besides ratings, review texts also contain a lot of information. The recent work [19] 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 users' expertise [20], temporal dynamics [15], and so on. Considering these factors may contribute to a better model; we plan to do this in our future work.

REFERENCES

[1] Wikipedia. 2017. Internet Water Army. Retrived from https://en.wikipedia.org/wiki/Internet_Water_Army.

[2] Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. 2014. De-biasing user preference ratings in recommender systems. In *Proceedings of the IntRS Workshop@RecSys'14*. 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 *Proceedings of the IntRS Workshop@RecSys'16*. 14–21.

[4] Mohammad Aliannejadi and Fabio Crestani. 2018. Personalized context-aware point of interest recommendation. *ACM Trans. Inf. Syst.* 36, 4, Article 45 (Oct. 2018), 28 pages. DOI: https://doi.org/10.1145/3231933

[5] Rolph E. Anderson. 1973. Consumer dissatisfaction: The effect of disconfirmed expectancy on perceived product performance. *J. Mark. Res.* 10, 1 (1973), 38–44.

[6] Jia Chen, Qin Jin, Shiwan Zhao, Shenghua Bao, Li Zhang, Zhong Su, and Yong Yu. 2016. Boosting recommendation in unexplored categories by user price preference. *ACM Trans. Inf. Syst.* 35, 2, Article 12 (Oct. 2016), 27 pages. DOI: https://doi.org/10.1145/2978579

[7] 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.

[8] 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.

[9] 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 4th ACM Conference on Recommender Systems*. ACM, 293–296.

[10] 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.

[11] Yong Ge and Jingjing Li. 2015. Measure and mitigate the dimensional bias in online reviews and ratings. In *Proceedings of the 36th International Conference on Information Systems, Fort Worth, TX*.
[12] 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.
[13] Radu Jurca, Florent Garcin, Arjun Talwar, and Boi Faltings. 2010. Reporting incentives and biases in online review forums. *ACM Trans. Web* 4, 2 (2010), 5.
[14] 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.
[15] 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.
[16] 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.
[17] Bibb Latané. 1981. The psychology of social impact. *Amer. Psychol.* 36, 4 (1981), 343.
[18] 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.
[19] 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.
[20] Julian McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd International World Wide Web Conference (WWW'13)*.
[21] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.
[22] Lev Muchnik, Sinan Aral, and Sean J. Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.
[23] Hung T. Nguyen, Preetam Ghosh, Michael L. Mayo, and Thang N. Dinh. 2017. Social influence spectrum at scale: Near-optimal solutions for multiple budgets at once. *ACM Trans. Inf. Syst.* 36, 2 (2017), 14.
[24] Richard L. Oliver. 2014. *Satisfaction: A Behavioral Perspective on the Consumer*. Routledge.
[25] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. *Introduction to Recommender Systems Handbook*. Springer.
[26] 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.
[27] Franklin E. Satterthwaite. 1946. An approximate distribution of estimates of variance components. *Biomet. Bull.* 2, 6 (1946), 110–114.
[28] 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.
[29] James Surowiecki, Mark P. Silverman, et al. 2007. The wisdom of crowds. *Amer. J. Phys.* 75, 2 (2007), 190–192.
[30] 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.
[31] Jian Wang and Yi Zhang. 2013. Opportunity models for e-commerce recommendation: Right product, right time. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*.
[32] 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.
[33] Larry Wasserman. 2013. *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media.
[34] 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.
[35] Fang Wu and Bernardo A. Huberman. 2010. Opinion formation under costly expression. *ACM Trans. Intell. Syst. Technol.* 1, 1 (2010), 5.
[36] Ming Yan, Jitao Sang, and Changsheng Xu. 2015. Unified YouTube video recommendation via cross-network collaboration. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR'15)*.
[37] Huiling Zhang, Md Abdul Alim, Xiang Li, My T. Thai, and Hien T. Nguyen. 2016. Misinformation in online social networks: Detect them all with a limited budget. *ACM Trans. Inf. Syst.* 34, 3 (2016), 18.
[38] Haiyi Zhu and Bernardo A. Huberman. 2014. To switch or not to switch: Understanding social influence in online choices. *Amer. Behav. Sci.* 58, 10 (2014), 1329–1344.

Received December 2018; revised July 2019; accepted August 2019