Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

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

Xiaoying Zhang, Hong Xie, Junzhou Zhao, John C.S. Lui
The Chinese University of Hong Kong
{jingyuxy, hongx87, junzhouzhao}@gmail.com, 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].

[1]This paper is an abridged version of the paper that won a best-paper award at the "Recsys2017" conference.

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 [McAuley *et al.*, 2015; Wang and Zhang, 2013] to online video/news platforms [Yan *et al.*, 2015; Davidson *et al.*, 2010]. 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 [Muchnik *et al.*, 2013; Salganik *et al.*, 2006; Adomavicius *et al.*, 2016; Weninger *et al.*, 2015] 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 [Muchnik *et al.*, 2013; Salganik *et al.*, 2006; Weninger *et al.*, 2015]. 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 [Adomavicius *et al.*, 2016]. As in [Adomavicius *et al.*, 2016], 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. [Wang *et al.*, 2014] 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 [Krishnan *et al.*, 2014; Adomavicius *et al.*, 2014] tried to mitigate the micro-level historical ratings' influence with an assump-

tion 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.

Contributions. Our contributions are as follows:

- **Observations.** By analyzing a dataset of 42 million ratings from *Tripadvisor* and *Amazon*, 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 depict 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 rating datasets from Amazon and TripAdvisor to study how historical ratings affect its next rating.

Formally, 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 $r_{p,i}$. And $e_{p,i} = \frac{1}{i-1}\sum_{k=1}^{i-1} r_{p,k}$ is called *prior expectation* of $r_{p,i}$.

Intuitively, Besides the influence from $\mathcal{H}_{p,i}$, there are two factors that could affect $r_{p,i}$: (1) the quality of product p; (2) user i's personal taste. We manage to control the last two factors, and plot the relationship between prior expectation and the average of the next rating [Zhang *et al.*, 2017]. Figure 1 shows the relationship in two selected groups in *Amazon-movie* dataset, and relationships in other groups and datasets are similar (presented in [Zhang *et al.*, 2017]).

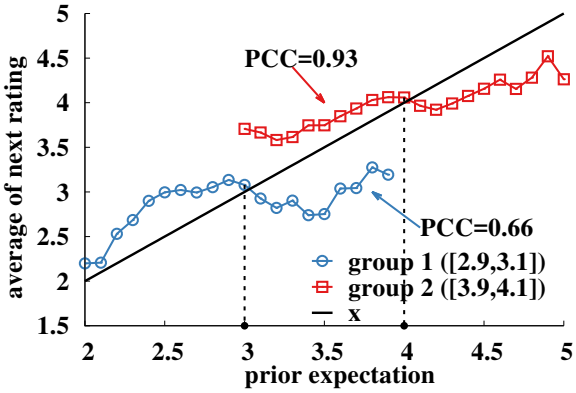


Figure 1: Relationship between prior expectation and the average of the next rating. The group 1 (group 2) contains products with average ratings in $[2.9, 3.1]$ ($[3.9, 4.1]$).

Observations. We obtain two observations from Figure 1:

- Products' historical ratings do affect the next rating. In general, the Pearson correlation coefficients (PCC) of relationships are in the range $[0.59, 0.94]$, this reflects a positive correlation between prior expectation and the next rating.
- Each curve with $\square$ ($\circ$) 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 1 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.

Explanations.

We next 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 [Oliver, 2014], 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.

For the second question, based on carefully designed hypothesis tests, we show that our observations can be well ex-

plained by the well-known *Assimilate-Contrast* theory [Anderson, 1973] in psychology [Zhang *et al.*, 2017].

- **"Assimilate-Contrast" theory**: If the disparity between his prior expectation and the product quality can be accepted by the user, 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.

Finally, we also verify that existing works on modeling historical ratings' influence [Krishnan *et al.*, 2014; Wang *et al.*, 2014; Adomavicius *et al.*, 2014] fail to 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.

3 Historical Influence Aware Latent Factor Model (HIALF)

In HIALF, the i-th rating of product p given by u is determined by three parts:

$$\hat{r}_{p,i,u} = b_u + q_{u,p} + \alpha_u h_{p,i} \quad (1)$$

- User u's personal bias b_u;
- The product p's quality in user u's view

  $$q_{u,p} = g + b_p + \boldsymbol{x}_u^T \boldsymbol{y}_p.$$

  Here g is the overall rating for an arbitrary user and product; b_p denotes item bias; $\boldsymbol{x_u}$ and $\boldsymbol{y_p}$ represent vectors of latent features for user u and product p.
- The distortion from historical ratings $\alpha_u h_{p,i}$, which is the product of how easily user u is affected by historical ratings (α_u) and the strength of historical influence ($h_{p,i}$). 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., $x = e_{p,i} - q_{u,p}$. Moreover, applying Latané's theory [Latané, 1981], 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

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

  More details on modeling $\boldsymbol{\beta}(x)$, $f(x)$, as well as a more general formula of $e_{p,i}$ (focus more on recent ratings) can be seen in our long paper [Zhang *et al.*, 2017]. We want to remark that we use a non-parametric way 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.

Finally, we use stochastic gradient descent (SGD) algorithm to infer the involved parameters by making the predicted ratings by HIALF as close as possible to real ratings.

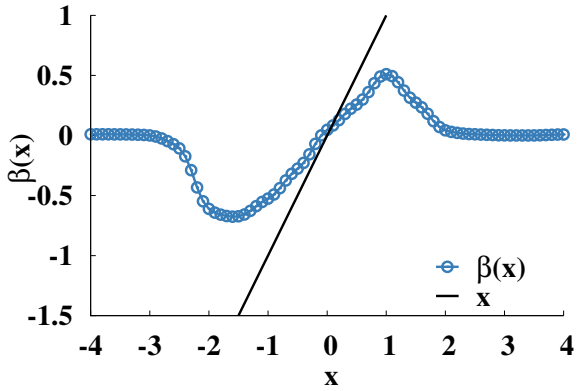


Figure 2: The learned $\boldsymbol{\beta}(x)$ on *Amazon-movie* dataset.

4 Experiments

We evaluate the performance of our model from three aspects: (1) whether the $\boldsymbol{\beta}(x)$ in HIALF meets with observations from real ratings, i.e., the *"Assimilate-Contrast"* theory; (2) how accurate HIALF can predict subsequent ratings compared to state-of-the-art models; (3) how well HIALF could fit the previous empirical observations in real ratings compared to state-of-the-art models.

Validating The Disconfirmation Bias Curve.

We find that all learned $\boldsymbol{\beta}(x)$ perfectly match the *"Assimilate-Contrast"* theory. Figure 2 shows the $\boldsymbol{\beta}(x)$ learned from *Amazon-moive* dataset. We can observe that in the range $[-2, 1]$, the bias assimilates to difference between prior expectation and the product quality, while deviating it out of the range. For $\boldsymbol{\beta}(x)$ learned form other datasets can be seen in our long paper [Zhang *et al.*, 2017].

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 [Wang *et al.*, 2014], latent factor (LF) model [Ricci *et al.*, 2011], 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.

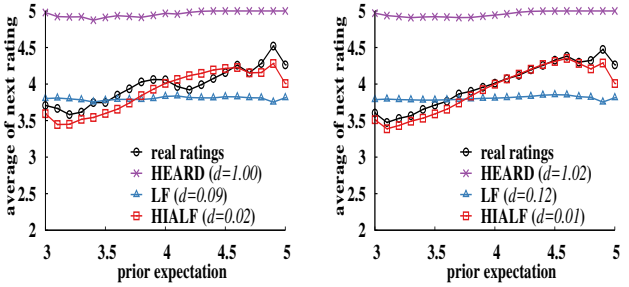
Table 1 shows that our model significantly outperforms alternatives on all datasets. On average, HIALF achieves a 33% reduction in MSE compared to HEARD, and a 6% reduction to LF. Furthermore, HIALF is consistently more accurate than HIALF-AVG, because users focus more on recent ratings when shaping prior expectations.

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. Detailed steps can be seen in our long paper [Zhang *et al.*, 2017]. Figure 3b and Figure 3a show the relationships for the product group with average ratings in $[3.9, 4.1]$ in *Amazon-movie*

	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 %

Table 1: MSE on five datasets



(a) prior expectation (average of historical ratings) (b) prior expectation (focus more about recent ratings)

Figure 3: Relationship between prior expectation and the average of the next rating in *Amazon-movie*. A smaller d implies a better fitting.

dataset. In Figure 3a, the prior expectation is the average of historical ratings, while in Figure 3b, prior expectation focuses more on recent ratings.

We can observe that HIALF provides the best fit to previous observations in real ratings. 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 fits the relationship of real ratings the best, as compared to LF and HEARD. We also define a quantative metric d to measure the difference between relationship in real ratings and in ratings generated by each model, and smaller d represents smaller difference. More details can be seen in the long paper [Zhang *et al.*, 2017]. 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.

5 Applications

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

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 in *debiased recsys*, the recommendation score for product p to user u is

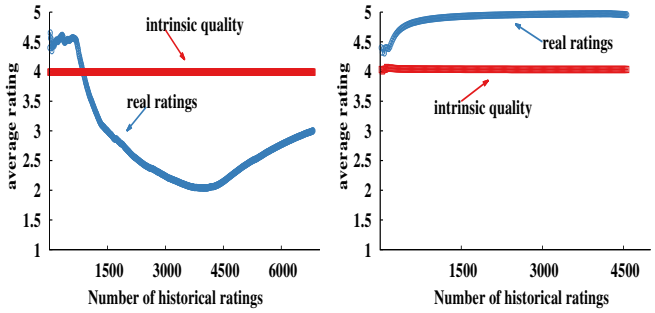
$$rec(p,u) = g + b_p + b_u + \boldsymbol{x_u}^T \boldsymbol{y_p}. \quad (2)$$

We then compare the *debiased recsys* with standard latent factor model (LF) on the set of ratings without historical ratings. The Root Mean Square Error (RMSE) is reported in Table 2. We can observe that *debiased recsys* consistently reveals smaller RMSE than LF, implying that it can provide better recommendations.

Exposing The Intrinsic Product Quality. The intrinsic quality of product p (q_p^*) is around the aggregated collective ratings given by a large group of users who were not exposed to historical ratings [Surowiecki *et al.*, 2007].

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**

Table 2: RMSE on five datasets



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

Figure 4: Two products with similar intrinsic quality have different rating growth histories, leading to significantly distinct ratings.

$$q_p^* = \frac{1}{N_p} \sum_{i=1}^{N_p} (g + b_p + \boldsymbol{x_{\tilde{u}(p,i)}}^T \boldsymbol{y_p}) \quad (3)$$

Here N_p is the number of ratings of product p, and $\tilde{u}(p,i)$ is the user who gave the i-th rating of product p. We use the case study in Figure 4 to illustrate the significance of revealing the intrinsic qualities of products. Figure 4 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, their average rating 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 Conclusion

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.

References

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

[Adomavicius *et al.*, 2016] Gediminas Adomavicius, Jesse Bockstedt, Shawn P. Curley, and Jingjing Zhang. Understanding effects of personalized vs. aggregate ratings on user preferences. In *IntRS Workshop@RecSys 2016*, pages 14–21, 2016.

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

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

[Krishnan *et al.*, 2014] Sanjay Krishnan, Jay Patel, Michael Franklin, and Ken Goldberg. 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*, pages 137–144, 2014.

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

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

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

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

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

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

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

[Wang and Zhang, 2013] Jian Wang and Yi Zhang. Opportunity models for e-commerce recommendation: Right product, right time. In *SIGIR*, 2013.

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

[Weninger *et al.*, 2015] Tim Weninger, Thomas James Johnston, and Maria Glenski. 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*, pages 293–297. ACM, 2015.

[Yan *et al.*, 2015] Ming Yan, Jitao Sang, and Changsheng Xu. Unified youtube video recommendation via cross-network collaboration. In *ICMR*, 2015.

[Zhang *et al.*, 2017] Xiaoying Zhang, Junzhou Zhao, and John Lui. Modeling the assimilation-contrast effects in online product rating systems: Debiasing and recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 98–106. ACM, 2017.