Spiral of Silence and Its Application in Recommender Systems

Chen Lin, *Member, IEEE*, Dugang Liu, Hanghang Tong, and Yanghua Xiao, *Member, IEEE*

Abstract—It is crucial to model missing ratings in recommender systems since user preferences learnt from only observed ratings are biased. One possible explanation for missing ratings is motivated by the spiral of silence theory. When the majority opinion is formed, a spiral process is triggered where users are more and more likely to show their ratings if they perceive that they are supported by the opinion climate. In this paper we first verify the existence of the spiral process in recommender systems by using a variety of different real-life datasets. We then study the characteristics of two key factors in the spiral process: opinion climate and the hardcore users who will give ratings even when they are minority opinion holders. Based on our empirical findings, we develop four variants to model missing ratings. They mimic different components of the spiral of silence based on the spiral process with global opinion climate, local opinion climate, hardcore users, relationships between hardcore users and items, respectively. We experimentally show that, the presented variants all outperform state-of-the-art recommendation models with missing rating components.

Index Terms—Spiral of silence, recommender system, missing not at random, opinion climate, hardcore

1 Introduction

Recommender systems, which discover items that users will be interested in is crucial to dissolve information overload problem in the age of consumption. Due to its high commercial value, the use of recommendation systems is no longer limited to E-commerce platforms, such as Amazon and Alibaba. As an indispensable module, recommendation function has been widely applied on the Internet, e.g., friend recommendation [1] on online social networking sites, transport recommendation in map Apps [2], news recommendation in mass media [3], citation recommendation in academic support systems [4] and so on. Therefore, recommender systems have received extensive attentions from both research communities and industries.

Recommender Systems fulfill their task by learning user preferences based on a collection of feedback. The major source of feedback is ratings which explicitly express user opinions. A number of approaches have been developed to learn user preferences from ratings, including memory based collaborative filtering methods [5], matrix factorization models [6] and its probabilistic expansions [7], and recently proposed neural networks [8]. The power of the aforementioned recommendation algorithms is highly dependent on the assumption that the collection of ratings correctly reflects the users' preferences. However, it is rare that users tell "the truth and the whole truth" at all times. On the contrary, ratings are often missing because the value of ratings affects users' willingness to give responses. In such cases, the representativeness of the ratings is degraded and the inference of a recommendation model is distorted.

Consider, for example, a toy data set illustrated in Table 1 where *Alice* is not willing to give her rating on the movie *Aliens*, because her true opinion is different from others. Suppose we adopt a userKNN recommendation [9] to predict *Alice*'s rating on movie *Eskiya*, we will consider userKNN with $k=1$, which is to find one nearest neighbor who shares the most similar taste with *Alice* and then take advantage of his/her ratings on *Eskiya*. In this toy example, the nearest neighbor should be *Diane*. However, as some ratings are missing, the system will make a wrong judgement that *Alice*'s nearest neighbour is *Bob*, based on the common item *Ben-Hur* that *Alice* and *Bob* have rated. Consequently, the predicted rating of *Alice* will be similar to *Bob*'s, which is 2 instead of 5.

Conventional matrix factorization models and their variants [10] will also give a biased prediction, as they are based only on observed ratings. A probabilistic explanation has been given in [11]. Without loss of generality, a matrix factorization model infers model parameters θ by maximizing the likelihood of observed ratings $p(R^{obs}|\theta)$. If the responses X (i.e., whether a rating is given) are dependent on the value of ratings R, then the true data likelihood (i.e., both observed ratings and responses) $p(X, R^{obs}|\theta) = p(X|R,\theta)p(R^{obs}|\theta)$ is not proportional to $p(R^{obs}|\theta)$. Therefore, parameters learnt from only observed ratings are not optimal.

Based on the above reasoning, it is crucial to take the non-random missing responses into account, i.e., incorporate $p(X|R,\theta)$ in the model. In the literature, MNAR models [11], [12], [13], [14], [15], [16] which assume ratings are

- Chen Lin and Dugang Liu are with the School of Informatics, Xiamen University, Xiamen 361000, China.
E-mail: chenlin@xmu.edu.cn, dgliu@stu.xmu.edu.cn.
- Hanghang Tong is with Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA. E-mail: htong@illinois.edu.
- Yanghua Xiao is with School of Computer Science, Fudan University, Shanghai 200433, China. E-mail: shawyh@fudan.edu.cn.

*Manuscript received 16 May 2019; revised 23 June 2020; accepted 31 July 2020. Date of publication 4 Aug. 2020; date of current version 29 Apr. 2022.
(Corresponding author: Chen Lin.)
Recommended for acceptance by K. Shim.
Digital Object Identifier no. 10.1109/TKDE.2020.3013973*

TABLE 1
Predict User Ratings on Eskiya, When Alice's Ratings on Aliens is Hidden

	Aliens	Ben-Hur	Casino	Dangal	Eskiya
Alice	*(2)*	3		3	5
Bob	5	3		2	2
Clare	5		5	1	2
Diane	2	2		3	5
Elle	5		2		2

Missing **N**ot **A**t **R**andom mimic the generation of responses $p(R|X,\theta)$. Although they have shown promising results, most previous MNAR models are based on relatively simple heuristics, e.g., the response is associated with the rating [12], [13], [14], or, the response is a function parameterized by the item feature [15], [16]. *Can we model the non-random missing ratings based on insights into user behavior that are supported by theoretical social studies?*

Multiple factors which can possibly cause and explain missing ratings have been studied in the literature, e.g., feedback loop [17], [18], [19] or selection bias in human decision making process [20], [21], [22]. However, the Spiral of Silence Theory [23], which is one of the most influential (as acknowledge in [24]) theories for explaining the formation and spread of opinion, has not been quantified, tested at scale or leveraged over recommender systems.[1] Fig. 1 illustrates the theory and its key factors: the *spiral process*, the perceived *opinion climate* and the *hardcore*. The theory states that, people are less willing to express their opinions if they perceive that they are not supported by the majority opinion. It results in a spiral process in which the majority opinion receives growing popularity while other opinions are gradually pushed back. The process reaches to a steady phase, when only the hardcore people remain to speak up for minority opinions and the majority opinion ultimately becomes a social norm. Though the theory is not the only reason that missing ratings exist in practice, it is suitable for modeling missing ratings in recommender systems since it provides theoretical connections between "opinion" (i.e., the rating values X) and "expression" (i.e., the response R) [25]. Thus, it is natural to model $p(X|R,\theta)$ under the theory framework. Furthermore, the theory distinguishes from other social theories such as "rich gets richer", "risk aversion", social conformity theory [26] and assimilation-contrast theory [27] since it provides counter-cases, i.e., hardcore people. Thus it is more flexible and generalizes well to the whole population.

We give an example from Amazon, the real-world recommender systems in Fig. 2 to explain the spiral process. The x-axis represents snapshots with increasing numbers of ratings. To reduce noise in the initial stage, we start the trend with at least ten ratings. The y-axis represents fraction of opinion holders. For visualization purpose, we divide $5-$star ratings into positive (4,5-star ratings) and negative opinions (1,2,3-star ratings) . In the early stages (less than 20 ratings), positive and negative opinions compete with each other. When positive opinions become majority opinions, we observe that the fraction of majority opinion holders gradually increases over time. Note that this is a spiral process while the small-scale fluctuations do not affect the overall upward tendency. Finally, the fraction of minority opinions does not approach zero, which means that despite of the dominant majority opinion, there is still a small number of hardcore users who persist their opinions.

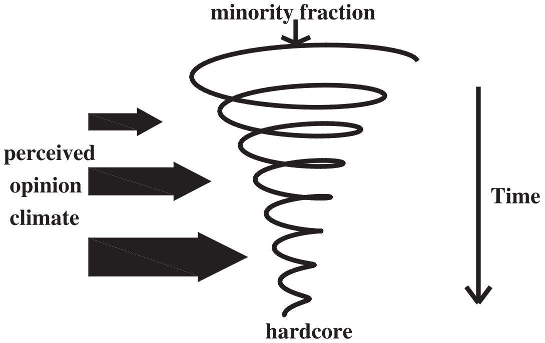


Fig. 1. Journey of minority fraction down the spiral of silence, which is induced by perceived opinion climate. The journey ends when only hardcores remain.

Our goal in this paper is not to provide the only possible explanation for missing ratings in recommender systems. Inspired by the spiral of silence theory, we (1) *provide a possible explanation that can coexist with others leading to non-random missing ratings in recommender systems*, and (2) *design better recommendation models based on the mechanism of non-random missing ratings*.

Toward these goals, we first empirically verify the existence of a spiral process (Section 2), i.e., users who perceive to accord with the majority opinion are more and more likely to show their ratings. Depending on how we model the opinion climate, we propose two MNAR models (Section 3).

We then study the properties of the hardcore users who are counter-cases of the spiral process (Section 4). The empirical findings of the personalities of hardcore users and their relationships with items lead to another two MNAR models (Section 5). We analyze the computational complexity of the proposed model variants (Section 6). We experimentally demonstrate the superior performance of the proposed four variants of MNAR models (Section 7).

The practical contribution of this paper is the empirical study which reveals the existence of spiral of silence in

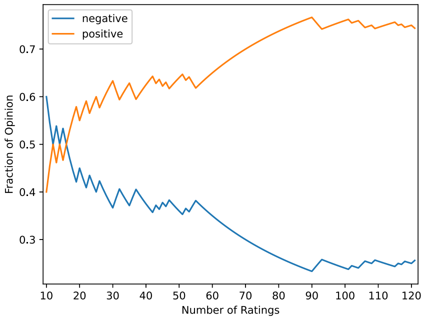


Fig. 2. Example: change of fractions of positive opinion holders (4,5-star ratings) and negative opinion holders (1,2,3-star ratings) of "The Preacher (Patrik Hedstrom and Erica Falck)", with different numbers of ratings.

[1] In the remaining of this paper, the Spiral of Silence theory will be referred to as "the theory".

recommender systems. Though the spiral of silence has been testified on many political issues [28], [29], [30], [31], their results are based on hypothetical willingness in lab environments [32]. On the contrary, we conduct large scale empirical study on real recommendation data sets with actual willingness. Our study is unique also because we emphasize on the time factor of the spiral process. We give formal definitions of the key factors in the theory, including opinion climate and hardcores, through quantitative study that can shed insights into niche marketing and efficient recommendations.

Moreover, the proposed models contribute to the recommendation community by integrating the effect of opinion climate towards missing ratings in MNAR models. We further explore the affect of other key factors of the spiral process, e.g., the hardcores, in recommender systems by a comprehensive comparative studies on four model variants.

2 Existence of Spiral of Silence

In this section, we testify the fundamental assumption of the spiral of silence theory in recommender systems: the existence of a spiral process. This section is a revised version of the empirical study in [33]. We focus on the scenario where only user-item ratings are provided, which is the most common setting in recommendation systems.

We first make the general assumption that users have a perceived public opinion in mind and have an idea of whether they are supported by the majority opinion. For example, on most rating sites, users are aware of the public opinion, e.g., previous ratings are displayed on the item's description page [34]. Note that a few exceptions may exist, for example, a user chooses to directly give low ratings after a bad consumption experience. Yet the exceptions do not affect the methodology we describe later.

We emphasize three important issues in verifying the existence of spiral process. First, the theory does not give a *formal definition of majority opinion*, especially for numerical ratings in recommender systems. Second, the perceived support indicates the *dominance of majority opinion*, i.e., majority opinion is strong enough to stifle debate. It is crucial to distinguish items depending on whether the spiral of process has been triggered. Third, the spiral process is dynamic and requires *trend analysis*. As time goes by, the majority opinion holders are more and more likely to show their ratings, resulting in an monotonically increasing fraction of majority opinion holders.

Therefore, in Section 2.1, we first study how to define the majority opinion. In Section 2.2, we explain the method to statistically testify the spiral process. We categorize items to two groups, i.e., with and without dominant majority opinion, and implement trend analysis on the fraction series of majority opinion holders. We report our results and verify the existence of spiral of silence in Section 2.3.

2.1 Majority Opinion

We use eight real data sets, including four Amazon product rating datasets [35], Epinions, Ciao [36], Movielens 20M [37] and Eachmovie [38]. All the ratings are timestamped. No other information is provided.

Suppose every user in the recommender system perceives the global opinion climate (i.e., same majority opinion for every user), let $\hat{r_{j,t}}$ denote the *majority opinion* of item j at time t. We consider five possible definitions of the majority opinion. The first two definitions are based on the average ratings on the particular item.

(1) Current average rating (current): $\hat{r_{j,t}} = \sum_{i,t'<t} r_{i,j,t'}/N(j,t)$, where $r_{i,j,t'}, t' < t$ is a rating given by any user i on item j before time t, $N(j,t)$ is the number of ratings given on item j before time t.

(2) Final average rating (final): $\hat{r_{j,t}} = \sum_{i,t'} r_{i,j,t'}/N(j)$, where $r_{i,j,t'}$ is a rating given by any user i on item j , $N(j)$ is the number of ratings given on item j. The above two definitions assume users have a statistical sense of item ratings.

The following four definitions assume a user likes to follow *opinion leaders*. Hence we have to select opinion leaders for each item, and then average their ratings, i.e., $\hat{r_{j,t}} = \sum_{i'} r_{i',j,t'}/N(j,i')$, where i' is an opinion leader, and $N(j,i')$ is the number of opinion leaders of item j. We propose four definitions to choose opinion leaders from the set of users who have rated the item.

(3) Active user average rating (active): the opinion leader are the most active users in the system, i.e., i' are the top 5 percent users with most ratings.

(4) Longest time-spanned user average rating (long): opinion leaders are the earliest and currently active users in the system. Since the datasets don't provide user registration time, we use the time range of the rating history instead, i.e., users i' are the top 5 percent users with the longest timespan between his first rating and last rating.

(5) Timely user average rating (timely): opinion leaders are the early adopters of this item who are eager to try new products and motivate other consumers, i.e., i' are the top 5 percent users with earliest ratings on this item.

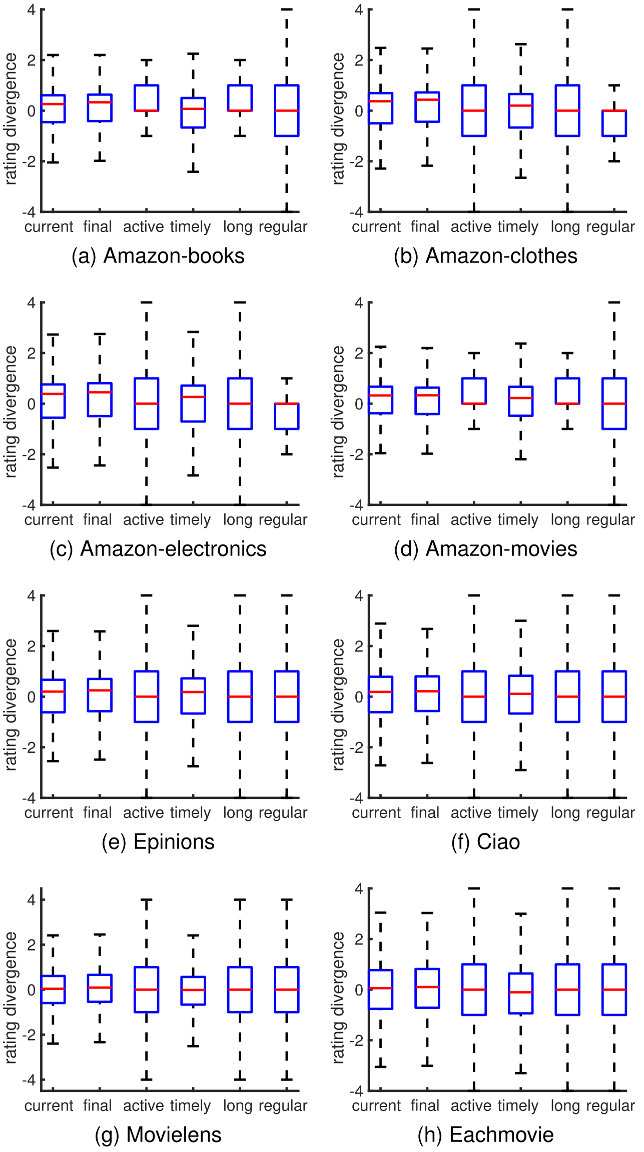
(6) Regular user average rating (regular): opinion leaders are regular users in the system, i.e., i' are the users who give at least one rating in every T months. Note that rating frequency in the e-commerce scenario may be lower than in other scenarios, we set $T = 2$ for Amazon datasets and $T = 1$ for other datasets.

We define the *rating divergence* $d(r_{i,j,t})$ of user i on item j at time t as:

$$d(r_{i,j,t}) = r_{i,j,t} - \hat{r_{j,t}}, \tag{1}$$

where $r_{i,j,t}$ is the rating by user i on item j given at timestamp t, $\hat{r_{j,t}}$ is the majority opinion at time t, as defined above.

We report the distributions of rating divergence in Fig. 3. The distributions are dataset specific. However, the current average rating and the final average rating have a relatively stable performance on all datasets. Similar conclusions can also be found in [39]. Another interesting finding is that, between regular user average rating and longest time-spanned user average rating, opinions of regular users are more useful for categories with higher consumption costs



(a) Amazon-books (b) Amazon-clothes

(c) Amazon-electronics (d) Amazon-movies

(e) Epinions (f) Ciao

(g) Movielens (h) Eachmovie

Fig. 3. Distribution of $d(r_{i,j,t})$ on eight data sets, based on six different definitions of majority opinion.

TABLE 2
Statistics of the Data Sets

Dataset	#Users	#Items	#Ratings
Amazon-books	8,026,324	2,330,066	22,507,155
Amazon-clothes	3,117,268	1,136,004	5,748,920
Amazon-electronics	4,201,696	476,002	7,824,482
Amazon-movies	2,088,620	200,941	4,607,047
Epinions	22,166	296,277	912,441
Ciao	7375	106,797	282,650
Movielens	138,493	131,262	20,000,263
Eachmovie	61,131	1622	2,558,871

(i.e., smaller divergence on clothes and electronics), and longest time-spanned users are more useful for categories with lower consumption costs (i.e., on books and movies).

2.2 Methodology

The first step is to filter items with a dominant majority opinion. Intuitively, if an item has a strong majority opinion, its ratings will be concentrated in a small range to form a peak. On the contrary, if an item does not have a strong majority opinion, the distribution of ratings will be flatter. Kurtosis is usually adopted to measure the level of consensus in social attitudes [40]. We use kurtosis to capture this information of each item, defined by:

$$k(j) = [E(r_j - \mu)^4]/[\sigma^4] - 3, \tag{2}$$

where random variable r_j is the rating of item j, μ is the mean of r_j, σ is the standard deviation of r_j, and $E(\cdot)$ is the expectation of a random variable. A normal distribution has kurtosis of 0. If the kurtosis is positive, the item has a dominant majority opinion. Otherwise if the kurtosis is negative, the item does not have a clear majority opinion.

The second step is to identify a time series for each item j, $< M_a(j)_s, \cdots, M_a(j)_e >$ from time s to time e, where each element in the sequence is the fraction of majority opinion holders. We start with quantifying the fraction of majority opinion holders, who give similar rating to the perceived majority opinion.

$$M_a(j)_t = \frac{|\{i|d(r_{i,j,t}) \in (-1,+1)\}|}{|N(j,t)|}. \tag{3}$$

It is not reasonable to include all snapshots, i.e., $M_a(j)_t, 1 \leq t \leq N$ in the sequence. Consider, for an extreme example, the majority opinion at the first time-stamp $r_{j,\hat{t}=1} = 1$ is really negative, and the majority opinion at the last time-stamp $r_{j,\hat{t}=N} = 5$ is quite positive. In this case, it is not fair to compare the majority opinion holders, because they are not the same group of people. Hence, we introduce *majority opinion expression* $M_e(j)_t$ associated with item j at time t. We define it as a floor function of the majority opinion $M_e(j)_t = \lfloor r_{\hat{j,t}} \rfloor$. We use the floor function because fluctuations in the convergence of average rating $r_{\hat{j,t}}$ is obviously limited in the range of $(-1,+1)$. For example, if $r_{\hat{j,t}}$ rose from 2.4 to 2.5, we see that the majority opinion does not change because the value of majority opinion expression remains $M_e(j)_t = 2$. For a period of time $s \leq t \leq e$, $M_e(j)_t$ is identical, then the sequence is associated with the item j, $s(j) =< M_a(j)_s, \ldots, M_a(j)_e >$.

Finally, we adopt the non-parametric Mann-Kendall (MK) test to detect monotonic trends in $s(j)$. The MK test compares each observation with its preceding observation and computes the following MK statistic $S(j)$ by

$$S(j) = \sum_{k=1}^{n-1} \sum_{i=k+1}^{n} sgn(M_a(j)_i - M_a(j)_k), \tag{4}$$

where sgn is a sign function and $M_a(j)_i$ is a time series element.

2.3 Results

The ratings in all the datasets are transferred into a 5-star scale in the experiment. We remove the items with less than 50 ratings in each dataset, because we need enough ratings to fully reflect the formation of opinions. We choose to use 10 ratings as a time window to segment time intervals.

In Table 3, we report the percentages of MK positive series, i.e., $|S(j) > 0, \forall j|/|S(j), \forall j|$ satisfying the different

TABLE 3
Percentage of Items (%) With $S(j) > 0$, at Different Significance Levels ρ

ρ	≤ 0.01		≤ 0.05		≤ 0.1	
Dataset	$k \geq 0$	$k < 0$	$k \geq 0$	$k < 0$	$k \geq 0$	$k < 0$
books	75.26	4.08	77.17	7.18	77.84	8.21
clothes	84.92	4.82	88.05	7.84	88.67	9.18
electronics	82.75	3.45	85.37	5.57	85.99	6.56
movies	77.26	4.73	79.99	7.12	80.95	8.52
Epinions	80.63	6.98	84.23	10.70	85.23	12.48
Ciao	74.38	6.41	76.35	14.10	77.83	17.95
Movielens	82.06	20.84	83.56	24.69	84.50	26.53
Eachmovie	68.20	16.17	68.20	20.80	70.29	22.14

significance levels for items with a dominant majority opinion ($k \geq 0$) and without a dominant majority opinion ($k < 0$). We can see that, no matter what the significance level we choose, for items with $k \geq 0$, most items are associated with a monotonically increasing time series of majority opinion holders. On the contrary, for items with $k < 0$, few items are associated with increasing majority opinion holders. This pattern holds for all data sets. Thus we verify the existence of a spiral process. For items with majority opinion, the proportion of majority opinion holders in population is monotonically increasing overtime until it reaches a stable status. When a spiral of silence is not triggered, the fraction of majority opinion holders will not increase. Note that the verification is from the perspective of majority opinion over the item universe. Though we can not exclude possible explanations from the perspective of weakening minority opinion, e.g., items simply attract a smaller audience, our study shows that the spiral of silence exists for a large portion of items.

Summary. In this section, we verify the existence of a spiral process in a variety of real recommender systems. We find that users whose ratings are similar to the majority opinion will be more likely to show ratings.

3 Spiral Process Model

In this section, we use the empirical findings in Section 2 to guide the developments of two recommendation models.

3.1 Preliminaries

For MNAR models, the observations in a recommender system include a set of ratings $R = \{r_{i,j}\}$, where a rating $r_{i,j}$ is given by user i to item j; and a set of responses $X = \{x_{i,j}\}$. If the user i has given a rating on item j, $x_{i,j} = 1$; otherwise, the user does not give a rating. More symbol notations are shown in Table 4.

TABLE 4
Notations

Variables	Explanations
Hyper-parameters	
σ	Variance for Gaussian distributions
ξ	Hyper-parameters for Beta distributions
Hidden-variables	
bv_j	Bias for item j
bu_i	Bias for user i
u_i	Preference vector for user i
v_j	Feature vector for item j
τ	Strength parameter
β	Hardcore persona probability
π_i	Binary persona variable for user i
Observations	
r_{ij}	Semi-observed rating on item j by user i
x_{ij}	Binary Response on item j by user i
e_j	Perceived opinion about item j
e_{ij}	Perceived local opinion climate before user i rates item j
c_i	Community indicator for user i
g_j	Group indicator for item j

For all the models presented in this paper, we assume that there are three distinctive stages when a user consumes an item in the recommender system: the *pre-rating stage*, the *rating stage* to generate a rating $r_{i,j}$, and the *post-rating stage* to generate a response $x_{i,j}$. The post-rating stages generate fully observable response, while the rating stages generate semi-observable ratings.

As with most matrix factorization models, we assume that there are K hidden aspects. The user preference is denoted as a vector $u_i \in \mathcal{R}^K$ for user i, and the item feature is denoted as a vector $v_j \in \mathcal{R}^K$ for item j. To encode additional information, scalars bu_i and bv_j denote user specific and item specific bias. The intuition of **P**robabilistic **M**atrix **F**actorization (PMF) [7] is that, a user will give a high rating if the item matches his/her preference. The rating $r_{i,j}$ approaches to $u_i v_j + bu_i + bv_j$, with a zero-mean Gaussian error, where u_i, v_j, bu_i, bv_j are all zero-mean Gaussian random variables. Therefore, in the pre-rating stage, the user preference u_i, user specific bias bu_i, item specific bias bv_j and item features v_j are generated from Gaussian distributions, $bu_i, bv_j \sim \mathcal{N}(0, \sigma_b^2)$, $u_i \sim \mathcal{N}(0, \sigma_u^2)$, $v_j \sim \mathcal{N}(0, \sigma_v^2)$. In the rating stage, the rating is generated from Gaussian distributions, $r_{i,j} \sim \mathcal{N}(u_i v_j + bu_i + bv_j, \sigma_r^2)$.

3.2 Global Opinion Climate Model MCO

As shown in Fig. 4, we model the spiral process where ratings are **M**issing **C**onditional on **O**pinion climate.

In the post-rating stage, model MCO assumes that, the user i has a perceived global opinion climate e_{ij} when rating item j, which in this model is the average rating of item j computed by Eq. (5). Note that we drop the subscripts i to reflect that parameters are tied across all i.

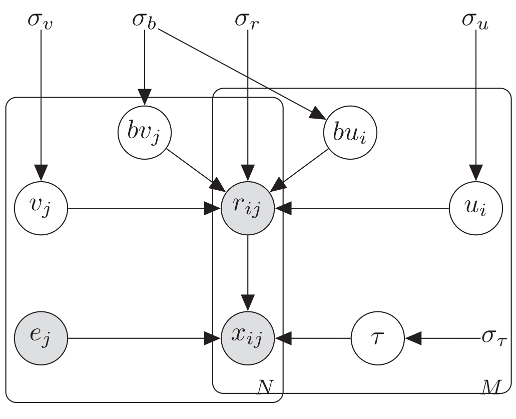


Fig. 4. Plate figure for global opinion climate model MCO.

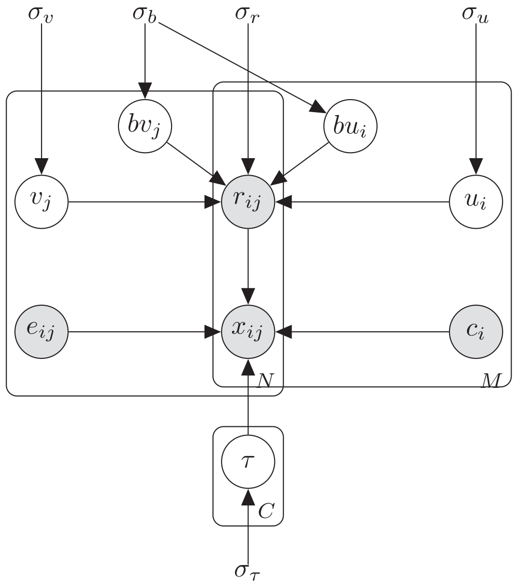


Fig. 5. Plate figure for local opinion climate model MCC.

$$e_j = \frac{\Sigma_i r_{i,j} x_{i,j}}{\Sigma_i x_{i,j}}. \tag{5}$$

Based on the empirical findings in Section 2, $x_{i,j} = 1$ has a higher probability if the rating divergence $|r_{i,j} - e_j|$ is small. Response $x_{i,j}$ is generated from Eq. (6):

$$P(x_{i,j} = 1 | r_{i,j}, e_j, \tau) = \frac{1}{\exp(\tau |r_{i,j} - e_j|)}, \tag{6}$$

where τ is a strength parameter, $\tau \sim \mathcal{N}(0, \sigma_\tau)$.

3.3 Local Opinion Climate Model MCC

Model MCO is based on global opinion climate, where each user observes the same majority opinion. One may argue that such a statistical sense of the global opinion climate is problematic. Rather, Internet users tend to have a limited vision of how others think and behave due to the infinite flow of information. They perceive local opinion climates based on communities they belong to. For example, many recommender systems provide social services, a user is frequently exposed to opinions formed by his friends, members in his interest groups, or similar users that hold opinions like his. For simplicity, we assume in this paper that a user belongs to one community. Note that in practice, a user may belong to multiple communities at the same time, and we leave it for future work.

Henceforth, we present model MCC, **M**issing **C**onditional on **C**ommunity opinion climate, which is shown in Fig. 5. Unlike MCO, MCC is based on local opinion climate, i.e., the majority opinion a user perceives is produced by his neighboring community. It is worthy to note that we attempted to embed the community generation process in the recommendation model. However, such a model leads to a result that we can not explain, i.e., best performance is achieved when the community number equals to two, with most users belonging to a large community. To preserve the explainability we externalize the community detection component, which will be described in Section 7.

Suppose we implement community detection method to partition the user space into C user groups. In the pre-rating stage, each user is assigned a community indicator c_i. Each community has a different strength parameter $\tau_c \sim \mathcal{N}(0, \sigma_\tau)$. The local opinion climate perceived by user i for item j is thus the average rating of all users in community c_i on item j, as computed by Eq. (7):

$$e_{i,j} = \frac{\Sigma_{c_{i'} = c_i} r_{i',j} x_{i',j}}{\Sigma_{c_{i'} = c_i} x_{i',j}}. \tag{7}$$

In the post-rating stage, the response $x_{i,j}$ is generated from Eq. (8):

$$P(x_{i,j} = 1 | r_{i,j}, e_{i,j}, c_i, \tau) = \frac{1}{\exp(\tau_{c_i} |r_{i,j} - e_{i,j}|)}. \tag{8}$$

4 Hardcore

Hardcores are counter-cases of the spiral process. Once majority opinion becomes powerful, minority opinion holders are pushed back, with only hardcore users left to display their ratings in defiance. Hardcore users are observed in any recommender system on every item. However, it remains an open question whether showing ratings (which are different from the majority opinion) is a random choice or a consistent behavior pattern for a user.

In this section, we first verify that hardcore personality exists, i.e., hardcore users in one recommender system are likely to behave hardcore in another recommender system. Then we analyze the correlation between hardcore users and items.

4.1 Hardcore Users

Our first question is whether hardcore is an inner character that shapes a user's behavior. We use the recent Yahoo! data set. The data set contains two sets of ratings: Yahoo! user and Yahoo!random. Yahoo!user set consists of ratings supplied by users during normal interactions, i.e., users pick and rate items as they wish. Yahoo!user resembles a "traditional" recommender system, which corresponds to a setting where users are free to hide their responses. Yahoo! random set consists of ratings collected during an online survey, when the same group of users in Yahoo!user set were asked to provide ratings on exactly ten items. Yahoo! random is different because the items are randomly selected by the system instead of the users themselves. Yahoo!random corresponds to a setting where users are forced to respond, against his actual willing. Note that, in constructing Yahoo!random dataset, "participants had the option of listening to a 30 second clip of each song"[2] to avoid noisy ratings when users rate music they don't have any opinion on. The dataset offers a unique opportunity to testify whether hardcore is a personality. If hardcore is an inner character, then the user will behave similarly under different settings. Therefore we present the following hypothesis.

Hypothesis 1 (H1). *Hardcore group in the user selected setting is similar to the hardcore group in the random setting.*

[2] https://webscope.sandbox.yahoo.com/catalog.php?datatype=r

TABLE 5
Statistics of the Data Sets Used in Section 4

Dataset	#users	#Items	#Ratings
Yahoo!user	15,400	1000	311,704
Yahoo!random	5400	1000	54,000

To testify H1, we first define hardcore group as a bunch of users who will give ratings no matter how the ratings diverge from the majority opinion. Based on the study in [33], we define N_i^h, which is the set of high divergent ratings of user i.

$$N_i^h = \{|r_{i,j,t} - \hat{r_{j,t}}| > \mu_2 + 0.5\sigma_2 = 1.7\}, \tag{9}$$

We compute a"hardcore" score h_i for each user i by Eq. (10).

$$h_i = |N_i^h \bigcap N_i|/|N_i|, \tag{10}$$

where N_i is the set of ratings a user i gives to all items.

We detect hardcore groups in both yahoo data sets with h_i exceeding a threshold $\tilde{h}$. Then compare the hardcore users in two subsets. For simplicity, we ignore the possibility that users use pseudonyms. We assume that each user is unique and represents one user in RS. To see whether the two hardcore groups are identical, we conduct Mann-Whitney U test to compare the overlap percentage between the two hardcore groups with a baseline overlap percentage given that users behave randomly (i.e., uniformly sample hardcore users from the two datasets). We find in Table 6 that, the two hardcore groups ($\tilde{h} \geq 0.5$) in different settings are identical, i.e., the overlap percentage of hardcore users is significantly larger than the baseline overlap. Furthermore, we discover that non-hardcore users ($\tilde{h} < 0.5$) are different under the two settings. Therefore H1 is verified. If a user is hardcore under one setting, he tends to be also hardcore under another setting.

4.2 Hardcore and Items

Another question is whether hardcore is related to moral basis. In the original theory [23], it is easy to understand that when there is a strong moral factor to the issue being debated, the minority opinion holders may more vociferously oppose the majority opinion, thereby create fierce controversy. However, hardcore moral in recommender system is still an unexplored problem.

We define two moral situations in Recommender Systems, one is to praise a (wrongly) criticized item (**PN**), the other is to criticize an (improperly) appreciated item (**CP**). Following the definition of hardcore score, we compute the percentage of high divergent ratings under two moral situations (1) **PN**: we compute $h_i^{PN} = |N_i^h \bigcap N_i^{PN}|/|N_i^{PN}|$ for each user i, where N_i^{PN} is the set of ratings that user i gives positive feedback (i.e., $r_{i,j} \geq 3$) to items with average negative feedback (i.e., $\hat{r} < 3$), N_i^h is defined by Eq. (9). Thus, h_i^{PN} describes user i's likelihood to give high divergent ratings (i.e., $r_{i,j} > \hat{r} + 1.7, r_{i,j} \geq 3, \hat{r} < 3$) to "save" a "bad" item. (2) **CP**: we compute $h_i^{CP} = |N_i^h \bigcap N_i^{CP}|/|N_i^{CP}|$ for each user i, where N_i^{CP} is the set of ratings that user i gives negative feedback (i.e., $r_{i,j} < 3$) to items with average positive feedback ($\hat{r} \geq 3$). Thus, h_i^{CP} depicts the user i's tendency to give high divergent ratings (i.e., $r_{i,j} < \hat{r} - 1.7, r_{i,j} < 3, \hat{r} \geq 3$) to criticize a "good" item.

TABLE 6
Percentage of Hardcore Group Overlap

Users	Non	Hardcore			
Threshold	$\tilde{h} < 0.5$	$\tilde{h} \geq 0.5$	$\tilde{h} \geq 0.6$	$\tilde{h} \geq 0.7$	$\tilde{h} \geq 0.8$
Actual	0.2016	0.1748**	0.0684**	0.0413**	0.0377**
Random	0.2272	0.1184	0.0304	0.0109	0.0075

** *indicates the actual overlap is significantly larger than random with significance level* $p \leq 0.05$ *based on Mann-Whitney U test.*

As shown in Fig. 6, in all datasets the percentage of high divergent ratings in CP (i.e., criticizing a positive item) is higher than the percentage in PN (i.e., praising a negative item). A possible underlying reason is that people feel more "obligated" to underrate a highly appreciated item than to save a criticized item. By analyzing the values of $|N_i^h|, |N_i^{PN}|, |N_i^{CP}|$ in Table 7, we can see that in most datasets, users generally prefer to give negative feedback on good items (i.e., larger $|N_i^{CP}|$ and smaller $|N_i^{PN}|$). The selection bias could also be one possible reason that leads to the phenomenon. As shown in Table 7, in each dataset, there are much more items which receive average positive feedback than items with average negative feedback (i.e., $P > N$).

Summary. In this section we verify that (1) hardcore is a personality with which users are likely to give deviant ratings in different settings; (2) users are more likely to give deviated ratings for items with overall positive feedback than for items with overall negative feedback.

5 Recommendation Models by Hardcore

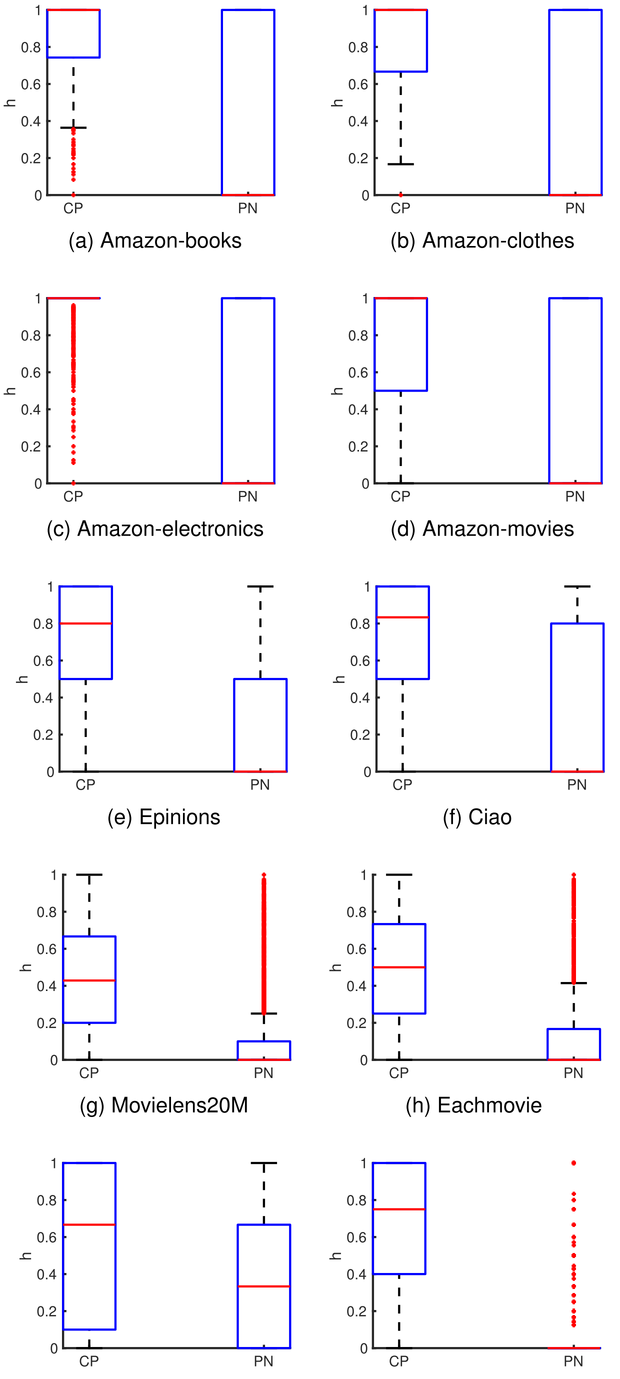
In this section, we develop two models based on the empirical findings in Section 4.

5.1 Hardcore User Model MCP

In model MCP, **M**issing **C**onditional on **P**ersona, we embed the personality of hardcore users. Hardcore users are more likely to give ratings that are not similar to the perceived opinion climate. Note that this model is consistent with the model proposed in [33].

As shown in Fig. 7, in the pre-rating stage, to model the split of users between hardcore and non-hardcore groups, we introduce a persona variable, denoted by $\pi_i \in \mathcal{R}^2$, an 1-of-2 coding for the persona indicator. The persona variable $\pi_i \sim Bern(\beta)$ is generated from a hardcore persona distribution. $\beta \in (0, 1)$ is generated from a Beta distribution $\beta \sim Beta(\xi_a, \xi_b)$ with hyper-parameters ξ_a, ξ_b. To model the behavior of hardcore and non-hardcore users, each persona is associated with a strength parameter $\tau_z \sim \mathcal{N}(0, \sigma_\tau), z \in \{0, 1\}$.

As verified in our empirical studies, the user is more likely to hide the rating if it is divergent to the perceived opinion climate. Furthermore, the user is more likely to display the rating if he/she is a hardcore user $\pi_{i,0} = 1$. These two findings together give us the following generation process in the post-rating stage:



(a) Amazon-books

(b) Amazon-clothes

(c) Amazon-electronics

(d) Amazon-movies

(e) Epinions

(f) Ciao

(g) Movielens20M

(h) Eachmovie

(i) Yahoo!user

(j) Yahoo!random

Fig. 6. Percentage of high divergent ratings under two moral situations. For PN, we compute h_i^{PN} for each user i, which is the fraction of i's high divergent ratings out of the set of ratings that user gives positive feedback to items with average negative feedback. For CP, h_i^{CP} is the fraction of high divergent ratings out of the set of ratings that user gives negative feedback to items with average positive feedback.

$$P(x_{ij}=1|r_{ij},e_{ij},\pi_i,\tau)=\prod_{z=0}^{z=1}\frac{1}{\exp(\tau_z|r_{ij}-e_{ij}|)^{\pi_{i,z}}}, \tag{11}$$

where e_{ij} is computed by Eq. (5).

5.2 Hardcore Item Model MCM

Our last empirical finding in Section 4 indicates that hardcore behavior is related to moral basis, i.e., different hardcore strength for items with positive feedback and negative feedback.

TABLE 7
Number of Items That Receive Negative Feedback N, Number of Items That Receive Positive Feedback P, Average Number of Ratings for $|N_i^h|$, $|N_i^{PN}|$ and $|N_i^{PN}|$ in Different Datasets

Dataset	N	P	mean($\|N_i^h\|$)	mean($\|N_i^{PN}\|$)	mean($\|N_i^{CP}\|$)
Amazon-books	557,074	21,950,081	0.0770	0.0183	0.0545
Amazon-clothes	275,175	5,473,745	0.0373	0.0163	0.0210
Amazon-electronics	464,815	7,359,667	0.0548	0.0261	0.0311
Amazon-movies	173,368	4,433,679	0.0765	0.0258	0.0592
Epinions	92,419	820,022	3.4501	1.4234	2.4870
Ciao	17,223	265,427	2.3429	0.6785	1.5440
Movielens	2,843,690	17,156,573	10.1836	11.3723	15.9924
Eachmovie	575,341	1,983,530	5.1109	5.2997	5.2570
Yahoo!user	172,433	139,271	6.9280	5.1154	2.8529
Yahoo!random	36,248	17,752	2.9380	1.3050	2.3270

Therefore, we present model MCM, **M**issing **C**onditional on **M**oral basis. The plate graph of model MCM is shown in Fig. 8. We group items on the moral basis, i.e., each item is assigned a group indicator $g_j \in \{0,1\}$. Items with an average rating greater than 3 belong to the positive group, and the remaining items form the negative group. Suppose, each group is associated with a different strength parameter $\tau_g \sim \mathcal{N}(0,\sigma_\tau)$.

In the post-rating stage, the response $x_{i,j}$ is generated from Eq. (12), where e_{ij} is computed by Eq. (5).

$$P(x_{i,j}=1|r_{i,j},e_{i,j},\tau)=\frac{1}{\exp(\tau_{g_j}|r_{i,j}-e_{i,j}|)}. \tag{12}$$

We apply Generalized EM algorithms to infer the parameters of model MCP and gradient descent methods to update parameters of model variants MCO, MCC and MCM.

6 Complexity Analysis

In this section, we analyze the time complexity of the proposed four models and compare them with traditional

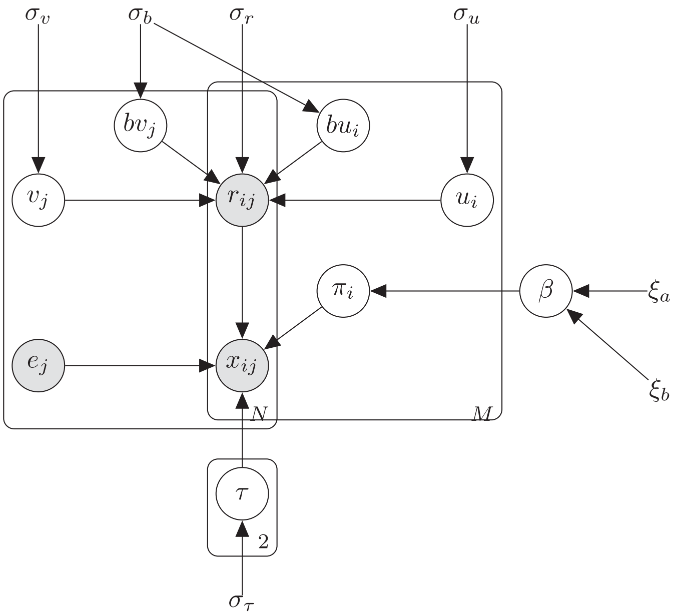


Fig. 7. Plate figure for hardcore user model MCP.

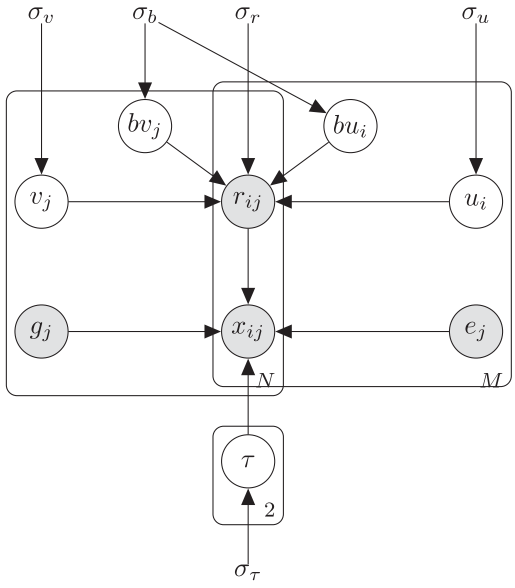


Fig. 8. Plate figure for hardcore item model MCM.

matrix factorization models and recommendation models with missing rating components.

Parameters of MCO, MCM and MCC are inferred via gradient descent algorithm. In each update, we need to calculate the gradient using the information of the entire matrix. Hence, their computation cost is $O(mnK)$, where m is the number of users, n is the number of items, and K is the number of dimensions. Note that in MCC, to perform the pre-clustering step, we need to equip an external community detection method as described in Section 7.3. Because the community detection method used is linear in the number of users $O(m)$, the total complexity of MCC is still $O(mnK)$. We apply the standard Expectation-Maximization algorithm to infer parameters of MCP. The E- and M-steps require computation over all items and users, with $O(mnK)$ cost.

The complexity of training the traditional matrix factorization model is $O(|\Omega|K)$ [7], where $|\Omega|$ is the size of observations. However, when the missing ratings need to be modeled, the complexity naturally extends from $O(|\Omega|K)$ to $O(mnK)$, such as RAPMF model [12], or even higher. For example, CPT-v and Logit-vd model [11] consider rating as multinomial variables, each update takes $O(nmKD)$, where D is the number of distinct rating values. A special case is the PropensityMF model [41] which keeps the complexity at $O(|\Omega|K)$.

Overall, in terms of complexity, the four methods we proposed are comparable to existing MNAR models. As training of recommender systems can be carried offline, it is worthy to spend more time on training if it can boost performance.

7 Experiment

In this section, we conduct comprehensive experiments to validate the performance of recommendation models motivated by the spiral of silence theory. Our aim is to answer three research questions.

1) Does modeling the spiral of silence process generate better recommendations?
2) Embedding which of the four factors can greatly boost the performance of a recommender system, global opinion climate, local community opinion climate, hardcore persona, or item-specific hardcore property?
3) How do the parameters affect the performance?

TABLE 8
Statistics of Data Sets in Experiments

Dataset	#users	#Items	#Ratings
Yahoo!user	15,400	1000	311,704
Yahoo!random	5400	1000	54,000
Coat Shopping user	290	300	6960
Coat Shopping random	290	300	4640

Codes of the proposed models are available at https://github.com/XMUDM/TKDE19-Spiral-of-Silence.

7.1 Experimental Setup

Datasets. To validate the model performance, we use two standard benchmark data sets in many MNAR recommendation model studies. Both of the data sets consist of test ratings that are randomly missing, i.e., users are “forced” to give ratings on randomly chosen items so that they can not hide their ratings. The Yahoo! data set has been used in the empirical study. It has been used to compare performances of almost all MNAR models [11], [12], [13], [14], [15], [16]. The coat shopping data set has been used in [41]. The models are trained on the “user” data sets, and tested on the “random” data sets. The statistics of the adopted datasets are listed in Table 8.

Evaluation Metric. The evaluation metric is NDCG@L, i.e., normalized discounted cumulative gain, which is a ranking performance measure that commonly adopted in evaluating recommender systems with missing data [11]. We report results up to $L = 10$ as in the Yahoo!random data, each user has only rated 10 items. Suppose an item at position j is associated with a relevance score $r(j)$, the NDCG score of a ranking system at top L results is computed by Eq. (13)

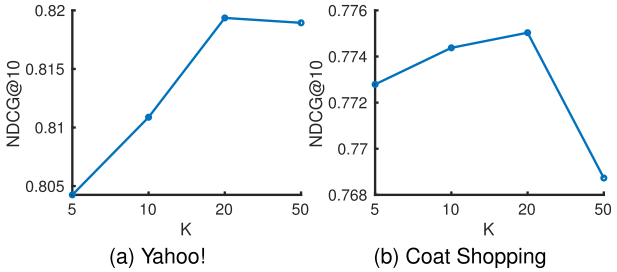
$$NDCG@L = \sum_{j=1}^{L} \frac{2^{r(j)} - 1}{\log(1+j)} / Z_L, \tag{13}$$

where $Z_L = \max NDCG@L$ is the normalization term, which is the maximal NDCG value obtained by the ground truth ranking list.

7.2 Effect of Parameters

In the following, we investigate how the parameters affect the performance of proposed model. We study the effect of standard deviation parameters $\sigma_u, \sigma_v, \sigma_b$ on u, v, bv, bu, σ_r for rating and the number of dimensions K. For simplicity, we set $\sigma_u = \sigma_v = \sigma_b = \sigma$. Due to the space limitation, we only report the results of $NDCG@10$ by using different values of K, σ and σ_r for the basic model MCO on both datasets. We observe similar results when L takes other values.

We first set $K = 5, 10, 20, 50$. Fig. 9 shows that as K increases, the performance of model MCO generally increases. When K gets too large, the performance is damaged, which is



(a) Yahoo! (b) Coat Shopping

Fig. 9. NDCG performance at top 10 items with varying values of K.

a phenomenon generally observed in matrix factorization models. However, the performance difference is not significant. The turning point is $K = 20$, which suggests that we should set K to be a relatively small number.

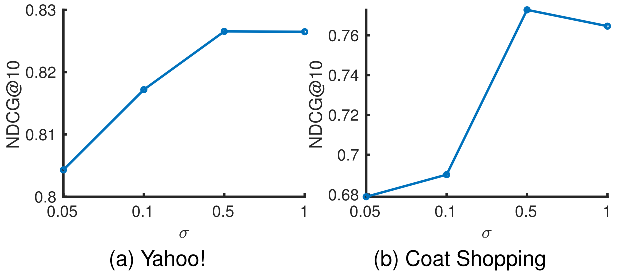
Varying the value of σ has a larger impact on NDCG performance than varying the value of K. We set $\sigma = 0.05, 0.1, 0.5, 1$. As shown in Fig. 10, for $\sigma < 0.5$, as σ increases, the $NDCG@10$ increases accordingly on both datasets. When $\sigma = 1.0$ the NDCG performance drops on Coat Shopping dataset. The decrease of NDCG performance when $\sigma = 1.0$ is insignificant on Yahoo! dataset. Similarly we set $\sigma_r = 0.05, 0.1, 0.5, 1$. As shown in Fig. 11, larger σ_r increases the $NDCG@10$ performance on both datasets. One possible explanation is based on the assumption of the spiral of silence models. With larger variance σ and σ_r, user ratings are more possible to deviate from the opinion climate, resulting in a larger probability of missing, i.e., smaller $p(x_{i,j=1})$ based on Eq. 6. Real-world recommender systems have very sparse data, i.e., most ratings are missing. Thus, higher σ and σ_r produce better performance.

7.3 Community Detection

The external community detection method for model MCC is implemented as follows. We first construct a behavior feature vector l_i for each user i. Intuitively, for a better recommendation performance, the community detection method must be able to distinguish user behavior patterns. Inspired by [42], the constructed user features is a three dimensional vector, i.e., $l_i = [n(i), pop(i), div(i)]$, where $n(i)$ is the number of ratings by user i, $pop(i)$ is the average popularity of items rated by user i calculated by Eq. (14), and $div(i)$ is the average rating divergence of user i calculated by Eq. (15).

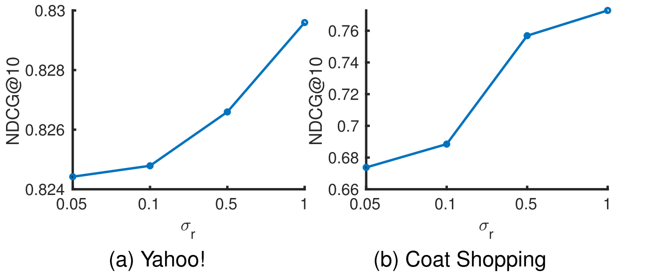
$$pop_i = \frac{\sum_{j \in v(i)} m(j)}{|v(i)|}, \tag{14}$$

where $m(j)$ is a function that gets the number of ratings for item j, $v(i)$ represents the set of rated item by user i.



(a) Yahoo! (b) Coat Shopping

Fig. 10. NDCG performance at top 10 items with varying σ values.



(a) Yahoo! (b) Coat Shopping

Fig. 11. NDCG performance at top 10 items with varying σ_r values.

$$div_i = \frac{\sum_{j \in v(i)} |r_{i,j} - \hat{r}_j|}{|v(i)|}, \tag{15}$$

where $\hat{r}_j$ is the current average rating of item j.

We then run K-means clustering on the behavior vectors. DBI (**D**avies-**B**ouldin **I**ndex) [43] is adopted to determine the number of clusters, i.e., the value of C. That is, we compute the inter-cluster distance S_c for cluster c

$$S_c = \left(\frac{1}{n(c)} \sum_{j=1}^{n(c)} |l_j - o_c|^2 \right)^{1/2}, \tag{16}$$

where $n(c)$ is the number of users in community c, o_c is the centroid vector of community c, which is the average vector of all users in this community.

We also compute the intra-cluster euclidean distance $M_{c,c'}$ for each pair of communities c, c'.

$$M_{c,c'} = \|o_c - o_{c'}\|_2 = \left(\sum_{k=1}^{n(o)} |o_{k,c} - o_{k,c'}|^2 \right)^{1/2}, \tag{17}$$

where $o_c, o_{c'} \in \mathcal{R}^{n(o)}$ are $n(o)$-dimensional centroid vectors, $o_{k,c}$ is the k-th component of the centroid vector o_c. For each cluster, we select the closest neighboring community as the index I_c of community c.

$$I_c \equiv \max_{c' \neq c} \frac{S_c + S_{c'}}{M_{c,c'}}. \tag{18}$$

The final index for C communities is averaged over all communities.

$$DBI(C) \equiv \frac{\sum_c I_c}{C}. \tag{19}$$

Finally, we vary the value of C and pick the value with the smallest $DBI(C)$. In the experiments below, the optimal number of community for Yahoo!random dataset is $C = 5$, for Coat Shopping dataset $C = 4$.

In Table 9, we compare the mean NDCG (i.e., average over $L = 1, 2, \ldots, 10$) results of the above mentioned method with traditional k-means clustering that only considers rating vectors and several graph-based community detection methods, such as the Louvain method [44], [45] and PCD method [46]. We can see that, K-means clustering with behavior vectors have generated the best performances.

TABLE 9
NDCG Result of MCC Model With Different Community Detection Methods

Method	Yahoo!	Coat Shopping
K-means+rating	0.6757	0.7487
Louvain	0.6794	0.7509
PCD	0.6802	0.7520
K-means+behavior	**0.7062**	**0.7826**

7.4 Comparative Study

Competitors. We compare our models to a wide range of available models, including conventional memory-based and model-based collaborative filtering recommenders and MNAR models. The competitors include (1) UKNN: the user based K-Nearest Neighbor collaborative filtering recommender [9]; (2) IKNN: the item based K-Nearest Neighbor collaborative filtering recommender [47]; (3) biasedMF: the matrix factorization model with user bias and item bias [6]; (4) PMF: the probabilistic matrix factorization model [7]; (5) CPT-v and (6) Logit-vd: both from the first MNAR models [11]; (7) PropensityMF: the principled approach to handle selection biases and adapt matrix factorization models [41] (8) RAPMF [12]: incorporates users' response models into the probabilistic matrix factorization. The parameters (including number of aspects K and variance σ) for the above models are tuned by cross validation.

The four proposed model variants include (1) MCO : missing conditional on global opinion climate, with the majority opinion to be current average rating; (2) MCC: missing conditional on local opinion climate, with the majority opinion to be averaged over the user's community. User community are detected by a k-means clustering method described above; (3) MCP: missing conditional on hardcore persona, where each user is associated with a persona variable. This is the same as the model proposed in [33]; (4) MCM: missing conditional on moral, where items are grouped into items with positive feedback and negative feedback, and the hardcore strengths are different for these two groups.

The default learning rate starts from 5×10^{-8} for u, v, bu, bv and 10^{-8} for τ. In each iteration, we increase or decrease the learning rate by likelihood comparison. Convergence is determined after a maximal number of 1500 rounds. The default hyper parameters for our proposed models are $K = 5, \xi_a = \xi_b = 2, \sigma = 0.5, \sigma_r = \sigma_\tau = 1$ for all Gaussian variances. The MCC model is preprocessed with K-means clustering with behavior vectors.

The comparative results are shown in Tables 10 and 11. We have the following observations. (1) All the proposed model variants based on spiral of silence outperform state-of-the-art competitors. They perform consistently and significantly better than all competitors in all NDCGs, on both Yahoo!random and coat shopping data sets. This shows the power of embedding spiral of silence theory in MNAR models. The spiral of silence theory does not only explain how ratings are missing in recommender systems, but also helps to design better recommendation models. (2) The best variant is MCC, which is to model local opinion climate by averaging ratings over the user's community. MCC performs consistently best in terms of all NDCGs on both datasets. The superior performance of MCC strongly suggests that in future MNAR models, the community structure must be incorporated. (3) MNAR models generally outperform traditional matrix factorization models. The best performing competitor is PropensityMF. However, the worst performing variant MCP still boosts the performance of PropensityMF by about 10 percent. This result demonstrates the competency of our model. (4) Furthermore, it is worth-noting that the persona specific strength parameter learnt for MCP model $\tau_1 = 2.2$ for non-hardcore users and $\tau_0 = 1.3$ for hardcore users on Yahoo! data set and $\tau_1 = 1.9, \tau_0 = 0.6$ on Coat Shopping data set. The interpretation for this value is that, for the same rating that falls in the minority opinion with high divergent $|r_{ij} - e_{ij}|$, a hardcore user is more likely to display the rating than a non-hardcore user. This result is consistent with the empirical findings.

TABLE 10
Comparison Results on Yahoo!Random Dataset

Metric	NDCG@1	NDCG@5	NDCG@10
UserKNN	0.4944	0.6582	0.7962
ItemKNN	0.4997	0.6536	0.7951
biasedMF	0.5432	0.6955	0.8172
PMF	0.5235	0.6667	0.8036
CPT-v	0.5767	0.7093	0.8275
Logit-vd	0.5724	0.7022	0.8249
RAPMF	0.5567	0.6944	0.8189
PropensityMF	0.5972	0.7207	0.8352
MCO	0.6425	0.7714	0.8616
MCP	0.6353	0.7617	0.8565
MCM	0.6369	0.7614	0.8562
MCC	**0.6709**	**0.7908**	**0.8730**

Underlined result indicates the variant performs significantly better than the best of competitors with significance level $p \leq 0.01$ based on Student's t-test. The best performance is also boldfaced.

8 Related Work

In this section, we briefly introduce related work on missing ratings in recommender systems, MNAR models, and spiral of silence.

TABLE 11
Comparison Results on Coat Shopping Dataset

Metric	NDCG@1	NDCG@5	NDCG@10
UserKNN	0.5699	0.6383	0.7224
ItemKNN	0.5878	0.6337	0.7214
biasedMF	0.4874	0.6172	0.7129
PMF	0.4959	0.5701	0.6809
CPT-v	0.5218	0.6075	0.7073
Logit-vd	0.4908	0.5754	0.6817
RAPMF	0.5164	0.6096	0.7089
PropensityMF	0.6173	0.6836	0.7612
MCO	0.6331	0.6922	0.7728
MCP	0.6313	0.6928	0.7732
MCM	0.6323	0.6942	0.7742
MCC	**0.6546**	**0.6963**	**0.7752**

Underlined result indicates the variant performs significantly better than the best of competitors with significance level $p \leq 0.01$ based on Student's t-test. The best performance is also boldfaced.

8.1 Missing Ratings in Recommender System

Real-world recommender systems are constrained by partially observed user-item interactions where massive ratings are missing. Missing ratings could be caused by multiple factors, two of which have received considerable research attention, i.e., feedback loop and user decision making process.

A recommender system's decision influences feedback from users, which in turn influences the system's decision, thus creating a feedback loop. Such a feedback loop may cause phenomenons such as echo chamber, i.e., users exposure only to recommendations based on others like themselves [18], popularity bias, i.e., exposure gradually decreases over time for long-tail items [48], and filter bubbles, i.e., people isolated from a diversity of content [49], and so on. Various practical methods have been proposed to degenerate feedback loops [17], [19], [41].

The user decision making process can cause missing ratings and rating bias. For example, previous research has shown that users have a filtering operation on the set of items based on a comprehensive consideration of multiple factors [21], which leads to missing ratings on certain items. Other recent work showed the strength of conformity, i.e., users will give biased ratings in accordance with others [34], the herding effect, i.e., new ratings follow previous ratings [50], and the Assimilation-Contrast effect [51], i.e., users will give similar ratings 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).

8.2 MNAR Models

MNAR model are probabilistic models that mimic the non-random missing process of responses. Here, non-random missing means the missing probability of a rating is relevant to the (hidden) value of rating. Some models relate a missing to simply the value of a hidden rating. For example, the earliest MNAR models CPT-v [11] assumes the response is sampled from one of R Bernoulli distribution, where R is the number of different discrete ratings. Similarly, RAPMF [12] models a response as a Bernoulli distribution which is parameterized by the rating scores for the observed ratings while as a step function for the unobserved ratings. MF-MNAR [14] first models the generation of ratings, then models a response matrix to "mask" the ratings. The response is probabilistic function of the value of ratings.

Other models relate the missing responses to the item to be rated. For example, the earliest MNAR models Logit-vd [11] assumes the response is generated by a sigmoid function governed by a parameter which is generated. The hierarchical Poisson factorization [15] models the overall responses for each item as a Poisson variable. The model in [16] first introduces a variable to indicate user exposure for each item, then models the rating generation process.

Finally, some models mix the above factors. Missing data mechanism in [13] is modeled by a Boolean OR operation of three Bernoulli random variable, each of which is related to users, items, and rating values.

Instead of modeling the missing mechanism directly, recent work [41] learns unbiased performance from data with selection bias by adapting models and estimation techniques from causal inference. This means that model can train each user's true preferences on self-selected data and apply directly to MNAR settings. To evaluate recommendation systems with MNAR ratings, appropriate surrogate objective functions are presented in [52] and a folding metric is investigated in [53] to quantify the likelihood of producing incongruous recommendations.

We can see that no previous MNAR work has been focused on the opinion climate. Furthermore they are unable to explain the evolution of ecology and several phenomena in the recommender systems, e.g., a high rated item gets more praises. Our work aims to reveal these hidden patterns from a social science perspective, and thus serves as a guiding light for future MNAR models.

8.3 Spiral of Silence

The spiral of silence theory has been widely acknowledged as a fundamental theory to explain the formation and spread of public opinion. The theory has been empirically verified in many political domains. Previous empirical study adopted a "train test" type of experiments, i.e., the subjects are questioned about perceived opinion climate and their willingness to discuss with a stranger on a train about any topic. Most works [28], [29], [30], [31] observe a positive correlation between perceived opinion climate and willingness to rate.

However those results are based on hypothetical willingness. We believe that our work is the first to verify the spiral model in large scale real life recommender systems. Moreover, they only proved the "social conformity hypothesis" [25]. Emphasis on time in the formation of the spiral has not been reflected on the methodologies. On the contrary, we acknowledge the dynamic nature of the spiral model.

We would like to differentiate the spiral of silence theory with the *rich gets richer* (Matthew effect) cliche. The "rich gets richer" assumption generates a similar phenomena with the spiral process, i.e., a stronger and dominating majority opinion. However, it does not relate the response to the value of a rating. Thus it is not as beneficial in designing a MNAR recommendation model.

9 Conclusion

Recommender systems are based on personalized user tastes. However, users are not isolated. They are highly influenced by public opinions. In this paper we bring a social science perspective to explain the missing ratings in recommender systems. We verify that, in recommender systems, users will perceive the opinion climate. Users who are supported by the majority opinion will be more and more likely to show their ratings. We study the factors which contribute to the formation of the spiral of silence, i.e. the definition of majority opinion, the existence of hardcore users and the characteristics of a hardcore person.

To demonstrate the impact of our empirical findings, we use the findings to guide the developments of four MNAR recommendation models. We show that all the proposed models outperform state-of-the-art models with and without MNAR assumptions. We also experimentally show that it is most important to model the local opinion climate, as users look in their community to have the sense of majority opinion.

We believe that our work can inspire models that not only produce good recommendation results but also preserve explainability. For future work, it is in our interest to expand the models to multiple modality. Furthermore, with the recent advance in neural recommendation models, it will be promising to combine the missing mechanism with deep neural networks.

Acknowledgments

This paper was supported by National Key R&D Program of China under No.2017YFC0803700, No.2017YFC1201200, by the National NSFC (No.61472085, 61472335, 61972328), by Shanghai Municipal Science and Technology project (No.16511102102, No.16JC1420401), Shanghai STCSM's R&D Program under Grant (No.16JC1420400).

References

[1] S. Wan, Y. Lan, J. Guo, C. Fan, and X. Cheng, "Informational friend recommendation in social media," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2013, pp. 1045–1048.

[2] R. Verma, S. Ghosh, M. Saketh, N. Ganguly, B. Mitra, and S. Chakraborty, "Comfride: A smartphone based system for comfortable public transport recommendation," in *Proc. 12th ACM Conf. Recommender Syst.*, 2018, pp. 181–189.

[3] C. Lin, R. Xie, X. Guan, L. Li, and T. Li, "Personalized news recommendation via implicit social experts," *Inf. Sci.*, vol. 254, pp. 1–18, 2014.

[4] X. Cai, J. Han, and L. Yang, "Generative adversarial network based heterogeneous bibliographic network representation for personalized citation recommendation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5747–5754.

[5] H. Luo, C. Niu, R. Shen, and C. Ullrich, "A collaborative filtering framework based on both local user similarity and global user similarity," *Mach. Learn.*, vol. 72, no. 3, pp. 231–245, 2008.

[6] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[7] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2008, pp. 1257–1264.

[8] H.-J. Xue, X.-Y. Dai, J. Zhang, S. Huang, and J. Chen, "Deep matrix factorization models for recommender systems," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 3203–3209.

[9] C. C. Aggarwal, "Neighborhood-based collaborative filtering," in *Recommender Systems*. Berlin, Germany: Springer, 2016, pp. 29–70.

[10] S. Rendle, "Factorization machines with libFM," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 1–22, 2012.

[11] B. M. Marlin and R. S. Zemel, "Collaborative prediction and ranking with non-random missing data," in *Proc. 3rd ACM Conf. Recommender Syst.*, 2009, pp. 5–12.

[12] H. Yang, G. Ling, Y. Su, M. R. Lyu, and I. King, "Boosting response aware model-based collaborative filtering," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2064–2077, Aug. 2015.

[13] Y.-D. Kim and S. Choi, "Bayesian binomial mixture model for collaborative prediction with non-random missing data," in *Proc. 8th ACM Conf. Recommender Syst.*, 2014, pp. 201–208.

[14] J. M. Hernández-Lobato, N. Houlsby, and Z. Ghahramani, "Probabilistic matrix factorization with non-random missing data," in *Proc. 31st Int. Conf. Int. Conf. Mach. Learn.*, 2014, pp. 1512–1520.

[15] P. Gopalan, J. M. Hofman, and D. M. Blei, "Scalable recommendation with hierarchical poisson factorization," in *Proc. 31st Conf. Uncertainty Artif. Intell.*, 2015, pp. 326–335.

[16] D. Liang, L. Charlin, J. McInerney, and D. M. Blei, "Modeling user exposure in recommendation," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 951–961.

[17] A. Sinha, D. F. Gleich, and K. Ramani, "Deconvolving feedback loops in recommender systems," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3243–3251.

[18] R. Jiang, S. Chiappa, T. Lattimore, A. György, and P. Kohli, "Degenerate feedback loops in recommender systems," in *Proc. AAAI/ACM Conf. AI Ethics Soc.*, 2019, pp. 383–390.

[19] B. P. Knijnenburg, S. Sivakumar, and D. Wilkinson, "Recommender systems for self-actualization," in *Proc. 10th ACM Conf. Recommender Syst.*, 2016, pp. 11–14.

[20] L. Chen, M. De Gemmis, A. Felfernig, P. Lops, F. Ricci, and G. Semeraro, "Human decision making and recommender systems," *ACM Trans. Interactive Intell. Syst.*, vol. 3, no. 3, pp. 1–7, 2013.

[21] A. Jameson *et al.*, "Human decision making and recommender systems," in *Recommender Systems Handbook*, Berlin, Germany: Springer, 2015, pp. 611–648.

[22] J. Lafky, "Why do people rate? theory and evidence on online ratings," *Games Econ. Behav.*, vol. 87, pp. 554–570, 2014.

[23] E. Noelle-Neumann, "The spiral of silence a theory of public opinion," *J. Commun.*, vol. 24, no. 2, pp. 43–51, 1974.

[24] J. D. Kennamer, "Self-serving biases in perceiving the opinions of others: Implications for the spiral of silence," *Commun. Res.*, vol. 17, no. 3, pp. 393–404, 1990.

[25] J. Matthes, "Observing the "spiral" in the spiral of silence," *Int. J. Public Opinion Res.*, vol. 27, no. 2, pp. 155–176, 2015.

[26] R. B. Cialdini and N. J. Goldstein, "Social influence: Compliance and conformity," *Annu. Rev. Psychol.*, no. 55, pp. 591–621, 2004.

[27] R. E. Anderson, "Consumer dissatisfaction: The effect of disconfirmed expectancy on perceived product performance," *J. Marketing Res.*, vol. 10, no. 1, pp. 38–44, 1973.

[28] W. De Koster and D. Houtman, "'Stormfront is like a second home to me': On virtual community formation by right-wing extremists," *Inf., Commun. Soc.*, vol. 11, no. 8, pp. 1155–1176, 2008.

[29] E. Nekmat and W. J. Gonzenbach, "Multiple opinion climates in online forums: Role of website source reference and within-forum opinion congruency," *Journalism Mass Commun. Quart.*, vol. 90, no. 4, pp. 736–756, 2013.

[30] P. Porten-Che é and C. Eilders, "Spiral of silence online: How online communication affects opinion climate perception and opinion expression regarding the climate change debate," *Stud. Commun. Sci.*, vol. 15, no. 1, pp. 143–150, 2015.

[31] A. Schulz and P. Roessler, "The spiral of silence and the internet: Selection of online content and the perception of the public opinion climate in computer-mediated communication environments," *Int. J. Public Opinion Res.*, vol. 24, no. 3, pp. 346–367, 2012.

[32] C. J. Glynn, A. F. Hayes, and J. Shanahan, "Perceived support for one's opinions and willingness to speak out: A meta-analysis of survey studies on the "spiral of silence"," *Public Opinion Quart.*, vol. 61, pp. 452–463, 1997.

[33] D. Liu, C. Lin, Z. Zhang, Y. Xiao, and H. Tong, "Spiral of silence in recommender systems," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, 2019, pp. 222–230.

[34] Y. Liu, X. Cao, and Y. Yu, "Are you influenced by others when rating? Improve rating prediction by conformity modeling," in *Proc. 10th ACM Conf. Recommender Syst.*, 2016, pp. 269–272.

[35] J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2015, pp. 785–794.

[36] J. Tang, H. Gao, and H. Liu, "mtrust: Discerning multi-faceted trust in a connected world," in *Proc. 5th ACM Int. Conf. Web Search Data Mining*, 2012, pp. 93–102.

[37] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interactive Intell. Syst.*, vol. 5, no. 4, pp. 1–19, 2015.

[38] P. McJones, "Eachmovie collaborative filtering data set," *DEC Syst. Res. Center*, vol. 249, 1997, Art. no. 57.

[39] H. Xie, Y. Li, and J. C. Lui, "Understanding persuasion cascades in online product rating systems," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 5490–5497.

[40] P. Di Maggio, J. Evans, and B. Bryson, "Have american's social attitudes become more polarized?" *Amer. J. Sociol.*, vol. 102, no. 3, pp. 690–755, 1996.

[41] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, "Recommendations as treatments: Debiasing learning and evaluation," in *Proc. 33rd Int. Conf. Int. Conf. Mach. Learn.*, 2016, pp. 1670–1679.

[42] C. Haydar, A. Roussanaly, and A. Boyer, "Clustering users to explain recommender systems' performance fluctuation," in *Proc. 20th Int. Conf. Foundations Intell. Syst.*, 2012, pp. 357–366.

[43] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[44] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *J. Stat. Mech.: Theory Exp.*, vol. 2008, no. 10, 2008, Art. no. P10008.

[45] R. Parimi and D. Caragea, “Community detection on large graph datasets for recommender systems,” in *Proc. IEEE Int. Conf. Data Mining Workshop*, 2014, pp. 589–596.

[46] J.-C. Ying, B.-N. Shi, V. S. Tseng, H.-W. Tsai, K. H. Cheng, and S.-C. Lin, “Preference-aware community detection for item recommendation,” in *Proc. Conf. Technol. Appl. Artif. Intell.*, 2013, pp. 49–54.

[47] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-based collaborative filtering recommendation algorithms,” in *Proc. 10th Int. Conf. World Wide Web*, 2001, pp. 285–295.

[48] H. Steck, “Item popularity and recommendation accuracy,” in *Proc. 5th ACM Conf. Recommender Syst.*, 2011, pp. 125–132.

[49] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan, “Exploring the filter bubble: The effect of using recommender systems on content diversity,” in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 677–686.

[50] G. Lederrey and R. West, “When sheep shop: Measuring herding effects in product ratings with natural experiments,” in *Proc. World Wide Web Conf.*, 2018, pp. 793–802.

[51] X. Zhang, J. Zhao, and J. C. Lui, “Modeling the assimilation-contrast effects in online product rating systems: Debiasing and recommendations,” in *Proc. 11th ACM Conf. Recommender Syst.*, 2017, pp. 98–106.

[52] H. Steck, “Training and testing of recommender systems on data missing not at random,” in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2010, pp. 713–722.

[53] D. Xin, N. Mayoraz, H. Pham, K. Lakshmanan, and J. R. Anderson, “Folding: Why good models sometimes make spurious recommendations,” in *Proc. 11th ACM Conf. Recommender Syst.*, 2017, pp. 201–209.

Chen Lin (Member, IEEE) received the BEng degree and the PHD. degree both from Fudan University, China, in 2004 and 2010. She is currently an associate professor with the School of Informatics, Xiamen University, China and Technology, Xiamen. Her research interests include web mining and recommender systems.

Dugang Liu received the MS degree from Xiamen University, Xiamen, China, in 2019. He is currently working toward the PhD degree in the College of Computer Science and Software Engineering, Shenzhen University, China. His research interests include recommender systems and counterfactual machine learning.

Hanghang Tong is currently an associate professor at the Department of Computer Science at the University of Illinois at Urbana-Champaign. Before that he was an associate professor at the School of Computing, Informatics, and Decision Systems Engineering (CIDSE), Arizona State University. He received the MSc and PhD degrees from Carnegie Mellon University in 2008 and 2009, both in Machine Learning. His research interest includes large scale data mining for graphs and multimedia. He has received several awards, including IEEE ICDM Tao Li Award (2019), SDM/IBM Early Career Data Mining Research Award (2018), NSF CAREER Award (2017), ICDM 10-Year Highest Impact Paper Award (2015), four best paper awards (TUP’14, CIKM’12, SDM’08, ICDM’06), seven ‘bests of conference’, one best demo, honorable mention (SIGMOD’17), and one best demo candidate, second place (CIKM’17). He has published over 100 refereed articles. He is the editor-in-chief of SIGKDD Explorations (ACM), an action editor of Data Mining and Knowledge Discovery (Springer), and an associate editor of ACM Computing Surveys (ACM), Knowledge and Information Systems (Springer) and Neurocomputing Journal (Elsevier); and has served as a program committee member in multiple data mining, database and artificial intelligence venues (e.g., SIGKDD, SIGMOD, AAAI, WWW, CIKM, etc.).

Yanghua Xiao (Member, IEEE) received the PhD degree in software theory from Fudan University, Shanghai, China, in 2009. Currently he is a full professor of computer science with Fudan University. His research interest includes big data management and mining, graph database, and knowledge graph.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.