Spiral of Silence in Recommender Systems

Dugang Liu, Chen Lin
Department of Computer Science, Xiamen University
Xiamen, China

Zhilin Zhang
School of Computing Science, Simon Fraser University
Burnaby, Canada

Yanghua Xiao
School of Computer Science, Fudan University
Shanghai, China
Alibaba Group
Hangzhou, China

Hanghang Tong
School of Computing, Informatics and Decision Systems Engineering, Arizona State University
Tempe, U.S.A.

ABSTRACT

It has been established that, ratings are missing not at random in recommender systems. However, little research has been done to reveal how the ratings are missing. In this paper we present one possible explanation of the missing not at random phenomenon. We verify that, using a variety of different real-life datasets, there is a spiral process for a silent minority in recommender systems where (1) people whose opinions fall into the minority are less likely to give ratings than majority opinion holders; (2) as the majority opinion becomes more dominant, the rating possibility of a majority opinion holder is intensifying but the rating possibility of a minority opinion holder is shrinking; (3) only hardcore users remain to rate for minority opinions when the spiral achieves its steady state.

Our empirical findings are beneficial for future recommendation models. To demonstrate the impact of our empirical findings, we present a probabilistic model that mimics the generation process of spiral of silence. We experimentally show that, the presented model offers more accurate recommendations, compared with state-of-the-art recommendation models.

KEYWORDS

Spiral of Silence, Recommender System, Missing not at Random, Hardcore

ACM Reference Format:
Dugang Liu, Chen Lin, Zhilin Zhang, Yanghua Xiao, and Hanghang Tong. 2019. Spiral of Silence in Recommender Systems. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3289600.3291003

1 INTRODUCTION

Recommender **S**ystems (RS) have received extensive attentions from both research communities and industries. The power of an RS is highly dependent on the assumption that the collection of ratings correctly reflects the users' opinions. Most recommender systems, however, suffer from extremely sparse rating data. More challengingly, it is rare that users tell "the truth and the whole truth" at all times. When a missing rating occurs due to the user's choice of non-response, the representativeness of the ratings is degraded and the inference of a recommendation model is distorted. Consider a conventional collaborative filtering RS running on a toy example illustrated in Table 1. Suppose, for some reason, *Alice* is not willing to give her ratings on the movie *Aliens* and *Eskiya*. The RS will make a wrong judgement that *Alice*'s nearest neighbor is *Bob*, based on the two common items *Alice* and *Bob* have, while in fact *Diane* shares the most similar taste with *Alice*.

Table 1: A toy example of 5 users' ratings on 5 movies. Alice's ratings on Aliens and Eskiya are hidden.

	Aliens	Ben-Hur	Casino	Dangal	Eskiya
Alice	*(2)*	3		3	*(5)*
Bob	5	3		2	2
Clare	5	4	5	1	2
Diane	2	2		2	5
Elle	5		2	4	2

In the literature of RS, models [1–6] which assume ratings are **M**issing **N**ot **A**t **R**andom (MNAR models) are recognized to have a superior ranking performance. Existing MNAR models mimic the generation of responses under different heuristics, i.e. the possibility of a response is related to the exact value of the rating [2–4] or to an unknown feature of the item [5, 6]. Unfortunately, none of the heuristics is empirically verified on real datasets, or supported by theoretical social studies.

In real scenarios, there could be various factors that lead to missing responses. Our goal in this paper is to provide a possible explanation for missing ratings and identify the key factors underlying users' decision process of whether or not to rate an item in recommender systems.

Towards this goal, we first empirically examine the response patterns in recommender systems. We verify the existence of a **spiral process** in which users are more and more likely to rate if they perceive that they are supported by the opinion climate (i.e. the dominant opinion), while the minority opinion holders are more and more reticent. Such a spiral process can be explained by the Spiral of Silence Theory[1] [7], which has been acknowledged

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5940-5/19/02…$15.00
https://doi.org/10.1145/3289600.3291003

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

as "one of the most influential recent theories of public opinion formation" [8].

We highlight the unique characteristics of our empirical study. (1) We study the behavior of giving ratings, while previous empirical studies focus on the biases in ratings [11–13, 33]. (2) We study the dynamic aspect which has never been addressed by existing MNAR models [1–6]. For example, as time passes by, the domination of majority opinion grows, which results in the increased willingness to rate for majority opinion holders and the decayed willingness for the minority opinion holders.

Two challenges arise in the empirical study. Firstly, survey studies in a lab environment [11–14] is problematic, because the findings are based on hypothetical willingness instead of actual willingness [15]. A recent survey [16] shows that the hypothetical willingness is a poor indicator of actual willingness. To address this challenge, we form our empirical study on the basis of actual willingness. Two scenarios are included: (1) Scenario I considers that a user's willingness to display or hide his/her ratings is offered (Sec. 3); (2) Scenario II considers that a user's inclination to rate is not available (Sec. 4).

Secondly, there are counter cases according to the Spiral of Silence Theory, i.e. the minority opinion that remains at the end of the spiral is called the **hardcores**. Mixing hardcore and non-hardcore users can hurt the performance of a recommendation model as the two user groups behave differently. To tackle this challenge, we present formal definition to distinguish hardcore users and study the characteristics of hardcore users in Sec. 5.

In order to demonstrate the impact of our empirical study, we present a straightforward application of the two major findings, i.e. the existence of spiral and the existence of hardcore users. We develop a **M**issing **C**onditional on **P**ersona (MCP) model which mimics the generation process of ratings and responses. The probability of giving response is related to the perceived opinion climate, individual rating, and the persona of the user (i.e. hardcore user or non-hardcore user). Experiments show that the MCP model outperforms state-of-the-art recommendation models, including models with and without MNAR assumptions.

Our contributions are three-fold.

Practical contributions. We verify the existence of a spiral of silence by large-scale empirical studies on 8 real recommendation data sets. We validate the group of hardcore users and provide detailed insights into the personal traits of hardcore users. These findings are particularly useful in niche marketing.

Methodological contributions. We design solutions to conduct large-scale field research of the spiral of silence theory on recommendation systems. For example, we give formal definitions to the core concepts including majority opinion holders and hardcore users. In hypothesis testing, we conduct a series of trend studies to capture the dynamic aspect of the theory which has hardly been addressed by previous studies.

Model contributions. We present a MCP model based on the major findings of our empirical study. We experimentally show that the MCP model outperforms state-of-the-art recommendation models. The significant improved performance reveals the potential impact of our empirical findings.

The paper is organized as follows. Sec. 2 summarizes the related works. In Sec. 3 and Sec. 4, we testify the existence of a spiral of silence over different scenarios. Sec. 3 is devoted to empirical study in which a user's willingness to rate is available. Sec. 4 corresponds to conducting empirical study on recommender systems in which a user's willingness to rate is not available. In Sec. 5, we study the existence of hardcore users in the formation of the spiral of silence, and we reveal the characteristics of hardcore users. Sec. 6 presents the MCP model which is developed by embedding the empirical findings. Sec. 7 presents and analyzes the experimental results. Finally, Sec. 8 concludes our contributions and insights into the future work.

2 RELATED WORK

Biases in Recommender System is related to our research. Some researchers observed a trend of increasing average ratings [17–19]; others found that later ratings are on average lower [20]. Hu et.al observed a J-shaped distribution [21]. A recent work [22] showed the existence and strength of conformity. These works focused on the biases in ratings, while we study the biases in responses, i.e. minorities are less likely to give ratings. In addition, the explanations in previous studies are not appropriate for all recommender systems, i.e. the choice-supportive bias [18] only applies to recommender systems with reviews.

The **rich gets richer** (Matthew effect) cliche is another line of research we want to distinguish our work with. Though the "rich gets richer" assumption generates a similar phenomena to the spiral process, it does not explain the formation of public opinion as the spiral theory does. The Matthew effect suggests group dynamics. It is not suitable to derive a recommendation model because personalization is not retained.

MNAR Models in RS are aware that ratings are missing not at random. Probabilistic models were presented to relate a missing to various factors, e.g. the value of a hidden rating [1–4] or to the item to be rated [1, 5, 6]. As we mentioned above, they were based on heuristics that are neither empirically verified nor theoretically proven. 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.

Empirical Study on Spiral of Silence Theory has a long history. They adopted a "train test" type of experiments, i.e. the subjects are questioned about their willingness to discuss with a stranger on a train about any topic. Most works [11–14] observe a positive correlation between perceived opinion climate and willingness to rate, both of which are collected during the survey of train test. However the result is 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" [9]. 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.

3 EXISTENCE OF SPIRAL: SCENARIO I

In this section we testify the fundamental assumption of the theory in recommender systems: the spiraling process, in which two key activities repeatedly occur. (1) A user is prompted to show his

rating if he perceives a majority opinion similar to his own but is restrained to show his rating when he believes his opinion belongs to the minority. (2) Such a response pattern leads to an even stronger majority opinion, which in turn encourages more majority opinion holders to show .

We need to define the core concepts: the *majority opinion* of ratings denoted as M_a, the *minority opinion* denoted as M_i. We also need to quantify the possibility to show a rating in majority opinion (denoted as p) and the possibility to show a rating in minority opinion (denoted as q). We present two solutions to automatically identify M_a, M_i, p, q in Sec. 3.1, a threshold-based approach and a model-based approach.

The data set used in this section is the extended epinions data set (Epinions) [23]. The data set contains ratings with their timestamps in a 5-star range. The number of users, items and ratings are shown in Table 2. Specially, this data set includes the display status of each rating, i.e. whether the user has chosen to show or hide his rating. Thus this data set is convenient for computing p, q.

Table 2: Statistics of the data set in Sec 3

Dataset	#Users	#Items	#Ratings
Epinions	120,492	755,760	13,668,320

3.1 Methodology

To start with, we assume that the *majority opinion* M_a is a set of ratings which are similar to the perceived opinion climate, and the *minority opinion* M_i, a set consists of ratings which are significantly divergent from the perceived opinion climate. 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 at timestamp t, $\hat{r_{j,t}}$ is the perceived opinion climate at time t. The value of $\hat{r_{j,t}}$ is defined as the average rating on item j before time t.

Threshold-based Approach. Most previous studies assume the majority opinion is a group of ratings with smaller absolute values of rating divergence and the minority opinion is with higher absolute values of divergence. We use a percentile based threshold. Since divergence is symmetric, we order all the rating divergences on each item and select a range $[Q_s, Q_e]$ around $d(r_{i,j,t}) = 0$, where Q_s is the starting percentile, and Q_e is the ending percentile. We define the *majority opinion* $M_a = \{r_{i,j,t} | Q_s < d(r_{i,j,t}) < Q_e\}$ as a a set of ratings that are positioned between the percentile range $[Q_s, Q_e]$. The rest of ratings are assigned to *minority opinion* M_i. We experiment with different ranges. For example, if we choose the 25th percentile for Q_s and 75th percentile for Q_e then the value of each rating in the majority opinion is higher than the bottom 25% of all ratings and lower than the top 25% of all ratings.

Given M_a, M_i and V the set of ratings which are displayed, we compute the willingness p, q as:

$$p = (|M_a \cap V|)/(|M_a|), \tag{2}$$

$$q = (|M_i \cap V|)/(|M_i|). \tag{3}$$

Model Based Approach. Instead of determining directly the range of majority opinion, we also propose a Gaussian Mixture Model which assumes the majority opinion follows a Gaussian distribution. Consider the Epinions data set as a collection of samples in the form of $(d(r_{i,j,t}), z_{i,j,t})$, where $d(r_{i,j,t})$ is rating divergence and $z_{i,j,t}$ is the display status, i.e. $z_{i,j,t} = 1$ indicates the user chooses to display the rating, $z_{i,j,t} = 0$ indicates the user chooses to hide the rating. We introduce a latent variable $M_{i,j,t}$ to indicate whether the rating belongs to the majority opinion. If the user considers himself as a majority opinion holder, his willingness to show the rating is $p = P(z_{i,j,t} = 1 | M_{i,j,t} = 1)$. Otherwise $q = P(z_{i,j,t} = 1 | M_{i,j,t} = 0)$. Inspired by [24], we assume that the divergence for a majority opinion rating is normally distributed with mean 0 and standard deviation σ_1, $P(d(r_{i,j,t}) | M_{i,j,t} = 1) \sim \mathcal{N}(0, \sigma_1^2)$. The mean is set to be zero because intuitively a majority opinion holder is most likely to be consistent with the opinion climate. σ_1 decides the range of majority. We assume that the absolute value of rating divergence for a minority opinion holder is distributed normally with mean μ_2 and standard deviation σ_2, $P(d(r_{i,j,t}) | M_{i,j,t} = 0) \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Using the EM algorithm, we can infer parameters $p, q, \sigma_1, \mu_2, \sigma_2$ and derive the latent variable $M_{i,j,t}$.

Given the formal definition of M_a, M_i, p, q, we develop the following two hypotheses. The first hypothesis focuses on a static phenomenon (corresponds to key activity 1). The second hypothesis emphasizes on the dynamic nature in the formation of the spiral (corresponds to key activity 2).

Hypothesis 1 (H1). *A majority opinion holder has a larger possibility to show the rating than a minority opinion holder, i.e. $p > q$.*

Hypothesis 2 (H2). *As the majority opinion becomes more dominant, the tendency for a majority opinion holder to show the rating is on the rise, while the tendency for a minority opinion holder to show the rating is on the decline, i.e. p increases over time while q decreases.*

To testify hypothesis H1, we identify M_a, M_i on the whole Epinions dataset and compare q, p. To testify hypothesis H2 we construct K snapshots. For each item, we sort the ratings in chronological order and equally divide them into K disjoint sets. We identify M_a, M_i on the K snapshots and conduct trend analysis on p, q.

3.2 Results and Analysis

Table 3: Willingness to show rating for majority opinion is significantly larger than minority opinion $p > q$.

Methodology	p	q
Threshold-based	0.5772	0.3715
Model-based	0.6094	0.2616

Table 4: Increasing willingness to show rating for majority opinion (positive p change) and decreasing willingness to show rating for minority opinion (negative q change).

Threshold-based approach							
K=♯snapshot	4	5	6	7	8	9	10
P change	0.284	0.309	0.329	0.343	0.352	0.357	0.361
q change	-0.206	-0.199	-0.195	-0.178	-0.178	-0.170	-0.165
Model-based approach							
K=♯snapshot	4	5	6	7	8	9	10
P change	0.343	0.366	0.377	0.394	0.401	0.405	0.407
q change	-0.249	-0.251	-0.235	-0.235	-0.226	-0.223	-0.218

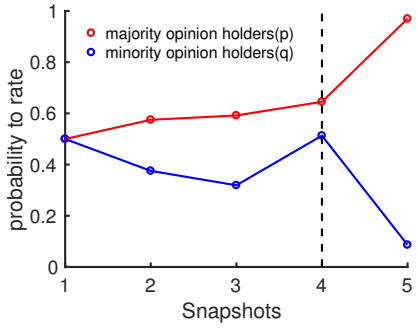


Figure 1: A case study: trend of p, q of item #498132 on Epinions, with 5 snapshots, by model-based approach.

As shown in Table 3, p is significantly larger than q, which means that the majority opinion holders are much more willing to show rating than minority opinion holders. $p > 0.5$ and $q < 0.5$ for both approaches, suggesting that this conclusion is not only in a comparative sense but also in an assertive sense. A majority opinion holder is likely to show rating ($p > 0.5$) and a minority opinion holder is prone to not showing ($q < 0.5$). Therefore H1 is verified. For the threshold-based methods, we find that reasonably expanding or narrowing the range does not change the values of p, q much. For ranges from ($Q_s = 25\% \sim Q_e = 75\%$) to ($Q_s = 15\% \sim Q_e = 85\%$) the values of p, q are in a narrow range $p \in (0.5757, 0.5773)$, $q \in (0.3630, 0.3731)$. For the model based approach, the solutions for parameters also give us some insights. Firstly, the mean parameter for the minority opinion is $\mu_2 = 1.3524$, $\mu_2 > \mu_1 = 0$, suggesting that the rating divergence for a minority opinion holder is much larger than that of a majority opinion holder. Secondly, the variance parameters are $\sigma_1 = 0.0744, \sigma_2 = 0.7521$. Note that in model-based approach, we do not explicitly require $\sigma_2 > \sigma_1$, hence, the result is reasonable because a minority opinion will be more divergent.

Fig. 1 illustrates the values of p, q at different snapshots for a typical item on Epinions. We can see that at the beginning the values of p, q are both close to 0.5, because majority opinion has not yet been formed. At this time both majority and minority opinion holders are active in discussion. p keeps rising until it reaches above 0.9 in the final snapshot, indicating that with a strong domination, most majority opinion holders are likely to give ratings. On the contrary q is shrinking to 0.35 in the third snapshot. We also notice that the fourth snapshot is a turning point. An interesting phenomena occurs in the forth snapshot, in which q slightly increases to 0.5. The minority opinion holders are arguing in defiance. We observe similar patterns in many other cases that the hardcores attempt to "save" the minority opinion before the turning point. Due to the limit of space, we do not give more case studies. After the turning point we observe steep slopes both in p rising from 0.6 to 0.95 and q declining from 0.45 to 0.05, which indicates that once majority opinion becomes powerful minority opinion holders are quickly pushed back and soon no alternative opinions exists.

For a detailed trend analysis, we report the difference between p, q in the last snapshot and p, q in the first snapshot in Table 4. We observe positive differences in p (p is larger in the last snapshot) and negative changes of q (q is smaller in the last snapshot). It means the tendency for a majority opinion holder to show rating is rising, while the tendency for a minority opinion holder to show rating is falling. H2 is verified. We manually check the values of p, q in each snapshot and find that p is monotonically increasing, which is an even stronger indicator that the spiral process exists. We do not observe monotone in q for all items, because of the revolt by hardcore people who are trying to "save" the minority opinion. Furthermore, we see that with more snapshots, the number of ratings in each snapshot is smaller, thus the trend is more significant, i.e. larger increase of p in the last snapshot.

Table 5: Statistics of the data sets in Sec 4 to Sec 5

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 s
Movielens20M	138,493	131,262	20,000,263
Eachmovie	61,131	1622	2,558,871

Summary and remarks. In scenario I where we know whether the user wants to hide or show his/her ratings, we verify the existence of a spiral of silence by showing (1) a majority opinion holder has a larger possibility to show rating than a minority opinion holder; (2) the possibility of a majority opinion holder to show rating is increasing as the majority opinion becomes more dominant.

4 EXISTENCE OF SPIRAL OF SILENCE: SCENARIO II

In most common settings, we can only obtain the ratings of recommender systems, without any other indicators that the user is willing to rate, i.e. whether the rating is hidden or displayed. Direct computation of willingness to rate is infeasible. To deal with recommender data sets without explicit willingness, we testify the existence of a spiral process by observing the trend of *majority opinion percentages*.

The majority opinion percentage is monotonically increasing if and only if there is a *majority opinion*. To see this we consider the consequence when a spiral of silence is triggered. If the majority opinion is strong, the majority opinion holders will be more active, resulting in an enhanced percentage of majority opinion. A majority opinion is necessary to trigger the silent spiral. When there is no clearly majority opinion, the opinion which is supported by more users than any other opinion is called a *plurality opinion*. A plurality opinion will not induce a spiraling process, thus the opinion percentages will not increase.

We use 8 real data sets, including four Amazon product ratings [25], product ratings datasets Epinions and Ciao [26] and movie rating datasets Movielens 20M [27] and Eachmovie [28]. All the ratings are timestamped. No other information is provided.

4.1 Methodology

The first problem that needs to be solved is to filter items with a majority opinion. Intuitively, if an item has a 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 majority opinion, the

distribution of ratings will be flatter. Kurtosis is usually adopted to measure the level of consensus in social attitudes [29]. We use kurtosis to capture this information of each item, defined by:

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

where random variable x_j is the rating of item j, μ is the mean of x_j, σ is the standard deviation of x_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 majority opinion. Otherwise if the kurtosis is negative, the item does not have a clear majority opinion (but it has a plurality opinion).

Given t the index of snapshots, $M_a(j)_t$ is the fraction of *majority opinion* (or *plurality opinion* for items without a majority opinion) at time t, and $M_e(j)_t$ is the associated *majority opinion expression* for item j at time t. Next, we present two strategies to compute $M_a(j)_t$ and $M_e(j)_t$.

Numerical Approach. The model based approach in Sec. 3 is not applicable in this scenario, as without the response variable it will be difficult to distinguish majority and minority. However it can be combined with the threshold-based approach to form a numerical approach. We first compute the rating divergence to the current average rating on the item for each rating as defined in Equ.(1). Note that in Sec. 3 we obtain $\sigma_2 = 0.75$, which suggests that the variance of the majority opinion is close to 1. Hence, to obtain the group of majority opinion holders, we define majority opinion on item j as a set of ratings $M_a(j)_t = \{i | d(r_{i,j,t}) \in (-1, +1)\}$. *Plurality opinion* is also calculated in the same way. *Majority opinion expression* is defined as a floor function of the average rating $M_e(j)_t = \lfloor \hat{r_{j,t}} \rfloor$. We use the floor function because fluctuations in the convergence of average rating $\hat{r_{j,t}}$ is obviously limited in the range of $(-1, +1)$. For example, if $\hat{r_{j,t}}$ rose from 2.4 to 2.5, we see that the majority opinion does not change because the value of majority opinion expression remains $M_e(j) = 2$.

Discrete Approach. Since the above method is based on a numerical estimation of opinion, one may argue that it is sometimes natural to represent an opinion by its polarity. Therefore in addition to the numerical representation of majority opinion, we present a discrete approach. We derive three segments, $SP = [3, 4, 5]$ are ratings for a positive opinion, $SE = [2, 3, 4]$ are ratings for a neutral opinion and $SN = [1, 2, 3]$ for negative opinions[2]. We allow the overlap of SP, SE, SN because such segmentation is more tolerant to different standards, i.e. some users will give a rating 2 to poor quality items while others will consider 2 a neutral opinion. We calculate the fraction of population on each segment and choose the highest segment as the *majority opinion* $M_a(j)_t = \{i | r_{i,j,t} \in S, S = \arg\max\{|SN|, |SE|, |SP|\}\}$. For example, if the percentages of population to rate 1, 2, 3, 4, 5 are 30%, 20%, 20%, 18%, 12% respectively, then the majority opinion is negative. We also calculate *plurality opinion* in the same way. The *majority opinion expression* $M_e(j)_t$ in discrete methods is expressed as negative, neutral or positive.

It is only meaningful to compare the majority opinion fractions $Ma(j)_t$ for different timestamp t under the same majority opinion expressions. Hence, for a period of time $s \leq t \leq e$, $M_e(j)_t$ is identical, then the sequence $< M_a(j)_s, \cdots, M_a(j)_e >$ is used in the following two hypotheses.

[2]Another commonly used segmentation, i.e. $SP = [4, 5]$, $SE = [3]$, $SN = [1, 2]$, also gives similar results to Table 7

Hypothesis 3 (H3). *For items with majority opinion, the proportion of majority opinion holders in population is monotonically increasing overtime until it reaches a stable status. Mathematically,* $\forall j, k(j) \geq 0, < M_a(j)_s, \cdots, M_a(j)_e >$ *is monotonically increasing.*

Hypothesis 4 (H4). *If the item has no clearly majority opinion, the proportion of its plurality opinion is unlikely to monotonically increase overtime. Mathematically,* $\forall j, k(j) < 0, ,< M_a(j)_s, \cdots, M_a(j)_e >$ *is not monotonically increasing.*

The non-parametric Mann-Kendall (MK) test is commonly employed to detect monotonic trends in time series data. The MK test compares each observation with its preceding observation and computes the following MK statistic (S) by

$$S = \sum_{k=1}^{n-1} \sum_{i=k+1}^{n} sgn(X_i - X_k), \tag{5}$$

where sgn is sign function and X is time series sample, i.e.$M_a(j)_n$. Note that we have to conduct MK test for each item.

4.2 Results and Analysis

Table 6: Percentage of items (%) with monotonically increasing $< M_a(j)_s, \cdots, M_a(j)_e >$ by numerical approach.

Significance level	≤ 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
Movielens20M	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

Table 7: Percentage of items (%) with monotonically increasing $< M_a(j)_s, \cdots, M_a(j)_e >$ by discrete approach.

Significance level	≤ 0.01		≤ 0.05		≤ 0.1	
Dataset	$k \geq 0$	$k < 0$	$k \geq 0$	$k < 0$	$k \geq 0$	$k < 0$
books	99.57	6.80	99.69	9.50	99.76	11.55
clothes	99.73	7.09	99.84	9.85	99.92	12.29
electronics	99.56	5.31	99.67	7.64	99.75	9.35
movies	99.42	8.70	99.62	11.78	99.76	13.96
Epinions	97.04	3.27	97.63	4.75	98.12	6.54
Ciao	98.24	12.82	98.24	15.38	98.82	16.67
Movielens20M	68.53	8.57	70.32	10.41	71.41	11.36
Eachmovie	93.31	10.09	94.14	11.23	94.56	12.26

The ratings in 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 6, we report the percentages of MK positive series $< M_a(j)_s, \cdots, M_a(j)_e >$ satisfying the different significance levels for items with a majority opinion ($k \geq 0$) and without a majority

opinion ($k < 0$), determined by the numerical approach. We can see that, no matter what the significance level we choose, for most items with a majority opinion ($k \geq 0$), the portion of majority opinion holders in population is increasing. On the contrary, for items without a majority opinion, very few of them (e.g. less than 4% at significance level $p \leq 0.01$) show a rising portion of plurality opinion. Thus the two hypotheses H3 and H4 are verified.

In Table 7, we report the MK positive series percentages determined by the discrete approach. The results are similar. Hypotheses H3 and H4 are again verified. The percentages obtained by the discrete approach are larger than that by the numerical approach. Because by discrete approach we will have a broader range for majority (or plurality) opinion, thus the possibility for observing a rising trend is bigger. For example, an item with a rating distribution centered on 2, 3 and 4 and its mean is 2.9, the majority opinion by discrete method is neutral (including 2, 3, 4), the majority opinion by numerical approach only includes ratings 2, 3.

Summary and remarks. In scenario II where no indicators of willingness are available, we verify the existence of a spiral of silence in recommender systems by showing (1) most items with a majority opinion have a monotonically increasing portion of majority opinion; (2) for items without a majority opinion, it is very unlikely that the proportion of its plurality opinion will monotonically increase.

5 FORMATION OF SPIRAL: HARDCORE

Hardcore is a key factor in the spiral of silence. In this section our objective is to understand the possible causes for users to act as hardcore.

5.1 Preliminaries

In RS, a hardcore group is a bunch of users who will give ratings no matter how the ratings diverge from the majority opinion. To define a hardcore user, we compute a"hardcore" score h,

$$h = n_i^h / n_i, \tag{6}$$

where n_i is the number of ratings a user i gives to all items, n_i^h is the number of high divergent ratings of user i. Using the results obtained in Sec. 3, the high divergent ratings are ratings $r_{i,j}$ where $\{|r_{i,j,t} - \hat{r_{j,t}}| > \mu_2 = 1.4\}$.

5.2 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. The dataset offers a unique opportunity to testify whether hardcore is a personality.

Table 8: Statistics of the data sets used in Sec 5

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

If hardcore is an inner character then the user will behave similarly under different settings. Therefore we present the following hypothesis.

Hypothesis 5 (H5). *Hardcore group in the user selected setting is similar to the hardcore group in the random setting.*

To testify H5, we first detect hardcore groups in both yahoo data sets by Eq.(6), and 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 9 that, the two hardcore groups ($h \geq 0.5$) 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 ($h < 0.5$) are different under the two settings. Therefore H5 is verified. If a user is hardcore under one setting, he tends to be also hardcore under another setting.

Table 9: Percentage of hardcore group overlap. ** indicates the actual overlap is significantly larger than baseline with significance level $p \leq 0.05$ based on Mann-Whitney U test.

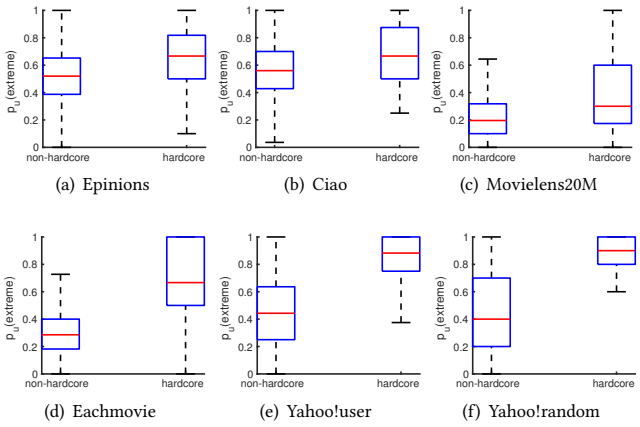
Users	Non-hardcore	Hardcore			
Threshold	$h < 0.5$	$h \geq 0.5$	$h \geq 0.6$	$h \geq 0.7$	$h \geq 0.8$
Actual	0.2016	0.1748^{**}	0.0684^{**}	0.0413^{**}	0.0377^{**}
Baseline	0.2272	0.1184	0.0304	0.0109	0.0075

Our next question is whether hardcore people are more likely to give extreme ratings. It is natural to relate hardcore with attitude certainty, while attitude certainty is represented by an extreme rating value. To study this, we set $h \geq 0.5$ to detect hardcore users, and plot the ratio of extreme ratings (i.e. ratings with values 1, 5) for hardcore and non-hardcore users in all data sets.

We can see from Fig. 2 that (1) in all data sets, hardcore users have a higher median ratio of extreme ratings. The IQR of $p_u(extreme)$ for hardcore users is higher than the IQR of $p_u(extreme)$ for non-hardcore users, which suggests that in recommender systems hardcore users are likely to give more extreme ratings. (2) Compare the two Yahoo!sets, we can see that hardcore users in Yahoo!random have a higher ratio of extreme ratings (i.e. smaller box and shorter tail). This observation is consistent to the spiral of silence theory, because when users are forced to rate (the Yahoo!random set), they can not hide extreme ratings (which will be missing in Yahoo!users as they are certainly different from the majority opinion).

5.3 Hardcore and Items

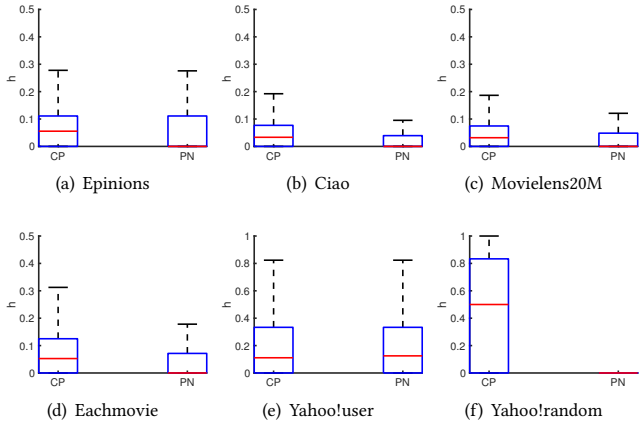
It is mentioned in [32] that hardcore is related to personal interest or importance. Some recommender systems encourage social tagging



(a) Epinions (b) Ciao (c) Movielens20M

(d) Eachmovie (e) Yahoo!user (f) Yahoo!random

Figure 2: Ratio of $1, 5$ ratings ($p_u(extreme)$) for hardcore and non-hardcore users in various datasets.



(a) Epinions (b) Ciao (c) Movielens20M

(d) Eachmovie (e) Yahoo!user (f) Yahoo!random

Figure 3: Hardcore score h under two moral situations, CP (criticize a positive item) and PN (praise a negative item).

to describe contents of items. In recommender systems with tags, personal interest is depicted by the number of ratings a user gives under a certain tag. Therefore we present the following hypothesis.

Hypothesis 6 (H6). *Under the most rated tag, an individual user has a higher hardcore score h.*

To testify this assumption, we use four data sets with tags, i.e. Epinions, Ciao, Movielens and Eachmovie. We count the number of ratings per user for each tag. For each user, we select the most rated tag and least rated tag (with at least 10 ratings to avoid bias). We compute the hardcore score in Equ. 6 for each user's most rated tag and least rated tag, where n_i is the number of ratings a user i gives to all items associated with the tag, n_i^h is the number of high divergent ratings i gives to items with the certain tag. We report the median h over all users in Table 10.

We can see that in all data sets, h is significantly higher under a most rated tag than a least rated tag. Thus H6 is verified. This is probably because for the most interesting items, users are more confident in their own experience and are more courageous to give a deviant opinion.

Another question is whether hardcore is related to moral basis. We define two moral situations in Recommender Systems, one is to praise a (wrongly) criticized item, the other is to criticize an (improperly) appreciated item. Following the definition of hardcore score, we compute h under two moral situations (1) CP, where n_i is the number of ratings that users give positive feedback ($r_{i,j} > \hat{r} + 1.4$) to items with average negative feedback ($\hat{r} < 3$), n_i^h is the number of high divergent ratings among n_i. (2) PN: n_i is the number of ratings that users give negative feedback ($r_{i,j} < \hat{r} - 1.4$) to items with average positive feedback ($\hat{r} \geq 3$), n_i^h is the number of high divergent ratings among n_i . As shown in Fig. 3, in most cases people feel more obligated to underrate a highly appreciated item than to save a criticized item.

Summary and remarks. In this section we study the hardcore factor in the formation of spiral of silence. We verify that (1) hardcore is a personality with which users are likely to give deviant ratings in different settings. (2) Hardcore is positively related to personal interest. We visualize that (1) hardcore users give more extreme ratings. (2) Users are more willing to criticize a (wrongly) appreciated item.

Table 10: The median hardcore score h for personally most rated tag and personally least rated tag. *** indicates significance level $p \leq 0.01$ based on Mann-Whitney U test.

Dataset	most rated tag	least rated tag
Epinions	0.1071^{***}	0
Ciao	0.0476^{***}	0
Movielens 20M	0.0625^{***}	0
Eachmovie	0.1250^{***}	0

6 MODEL

In this section, we develop a **M**issing **C**onditional on **P**ersona (MCP) model. We embed two major empirical findings in the MCP model. (1) Users are more likely to give ratings if their ratings are consistent with the perceived opinion climate. (2) Hardcore users are more likely to give ratings that are not similar to the perceived opinion climate.

6.1 Preliminaries

As with most matrix factorization models, we assume that there are K hidden aspects. The user preference is denoted as a vector $U_i \in R^K$ for user i, and the item feature is denoted as a vector $V_j \in R^K$ for item j. The rating given by user U_i to item V_j is denoted by $X_{i,j}$. The intuition of **P**robabilistic **M**atrix **F**actorization (PMF) [34] is that , a user will give a high rating if the item matches his/her preference. Therefore, the rating $X_{i,j}$ approaches to $U_iV_j + BU_i + BV_j$, with a zero-mean Gaussian error, $X_{i,j} \sim \mathcal{N}(U_iV_j + BU_i + BV_j, \sigma_r^2)$, where BU_i and BV_j are user specific and item specific bias. U_i, V_j, BU_i, BV_j are all zero-mean Gaussian random variables.

6.2 Modeling the Spiral of Silence

With respect to the user rating process, we assume that there are three distinctive stages: pre-rating stage, rating stage, and post-rating stage. As shown in Fig. 4, the MCP model mimics the following generation process.

The pre-rating stage. In this 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. $BU_i, BV_j \sim \mathcal{N}(0, \sigma_b^2)$, $U_i \sim \mathcal{N}(0, \sigma_u^2)$, $V_j \sim \mathcal{N}(0, \sigma_v^2)$.

To model the split of users between hardcore and non-hardcore groups, we introduce a persona variable, denoted by $\pi_i \in R^2$, 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 $\tau_z \sim \mathcal{N}(0, \sigma_\tau), z \in \{0, 1\}$.

The rating stage. Similar to PMF, user i generates a rating X_{ij} for item j based on his/her rating bias, the item's rating bias, his/her preferences and the item features:

$$X_{ij} \sim \mathcal{N}(U_i V_j + BU_i + BV_j, \sigma_x^2) \tag{7}$$

The ratings are semi-observed. Whether the rating X_{ij} is observed is denoted by a binary response variable $R_{i,j}$, where $R_{i,j} = 1$ indicates the rating is observed and otherwise the rating is missing.

The post-rating stage. In this stage, the user decides whether or not to reveal his/her rating. The user will first perceive the opinion climate E_{ij}, which in this model is an observed variable. E_{ij} is defined as the average rating on item j before i's rating if the rating X_{ij} is observed. If X_{ij} is not observed, we use the average rating on item j to approximate the opinion climate at the time of rating.

As verified in our empirical studies, (1) the user is more likely to hide the rating if it is divergent to the perceived opinion climate; (2) 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:

$$P(R_{ij} = 1 | X_{ij}, E_{ij}, \pi_i, \tau) = \prod_{z=0}^{z=1} \frac{1}{\exp\left(\tau_z |X_{ij} - E_{ij}|\right)]^{\pi_{i,z}}}. \tag{8}$$

We apply a Generalized EM algorithm to infer the model parameters $U, V, BU, BV, \beta, \tau$. In the E-step we estimate latent variable π_i for each user. In the M-step, we update the model parameters by gradient descent. The inference can be found in supplementary material.

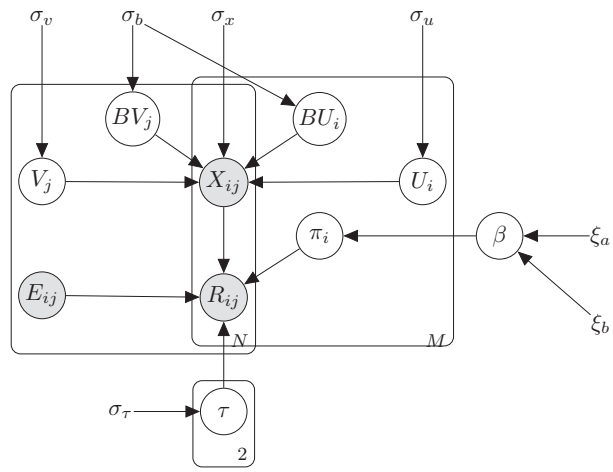


Figure 4: Plate graph of the proposed MCP model.

Table 11: Notations for MCP model

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
β	Hardcore persona probability
π_i	Binary persona variable for user i
Observations	
X_{ij}	Rating on j by i
R_{ij}	Binary Response on j by i
E_{ij}	Perceived opinion climate before R_{ij}

7 EXPERIMENT

We describe a brief experimental analysis of the MCP model in this section. More experimental results can be found in supplementary material.

The major evaluation metric is $NDCG@L$, which is a standard measure for ranking systems.

The comparative $NDCG$ is conducted on Yahoo!random dataset. Evaluating $NDCG@L$ on a randomly missing data set, such as Yahoo!random, has been used as the primary criteria in many MNAR researches [1, 4]. Experiments on other non-randomly missing datasets could be misleading as the ground truth does not accurately reflect user preferences. Therefore, $NDCG$ on Yahoo!random dataset is the primary evaluation metric in this work.

We compare our model to a wide range of available models, including conventional memory-based and model-based collaborative filtering recommenders and MNAR models. The comparative models include (1)UKNN: the user based K-Nearest Neighbor collaborative filtering recommender; (2) IKNN: the item based K-Nearest Neighbor collaborative filtering recommender; (3) MF: the standard matrix factorization model [35]; (4)PMF: the probabilistic matrix factorization model [34]; (5)CPT-v and (6) Logit-vd: both from the first MNAR models [1]; (7) MF-MNAR [4]: the recent probabilistic MNAR model which masks the rating matrix by a response matrix; (8)RAPMF [2]: a recent MNAR model which 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 hyper-parameters for MCP model is $K = 5, \xi_a = \xi_b = 2, \sigma = 0.5$ for all variances. Learning rate is initialized with $1e^{-8}$

and decayed every 10 rounds. Convergence is determined after a maximal number of 3000 rounds. The reported results are averaged over 5-fold validation.

We can see in Fig.5 that MCP performs consistently best in all NDCGs. It boosts the performance for about 5% than the best of MNAR models, i.e. MF-MNAR. This result demonstrates the competency of our model. Furthermore, it is worth-noting that the persona specific strength parameter learnt for MCP model $\tau_1 = 2$ for non-hardcore users and $\tau_0 = 0.4$ for hardcore users. The interpretation for this value is that, for the same rating that falls in the minority opinion with high divergent $|X_{ij} - E_{ij}|$, a hardcore user is more likely to display the rating than a non-hardcore user. This result is consistent with the empirical findings.

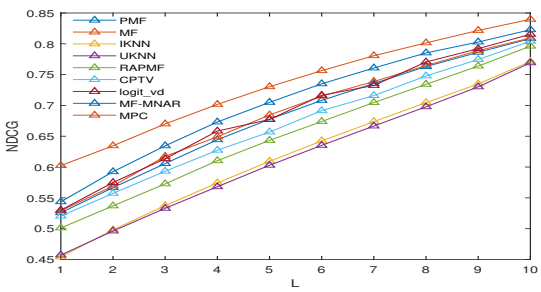


Figure 5: Comparable NDCG performance at top L items

8 CONCLUSION

In this paper we bring a social science perspective to the empirical study of missing not at random ratings in recommender systems. We verify the spiral of silence theory in large-scale real recommendation systems. We study the factors which contribute to the formation of the spiral of silence, i.e. the existence of hardcore users and the characteristics of a hardcore person. Our findings not only reveal that ratings in recommender systems are not missing at random, but also capture the mechanism of missing ratings. To demonstrate the impact of our empirical findings, we use the findings to guide the developments of a MNAR recommendation model. We experimentally show that such a model outperforms state-of-the-art models with and without MNAR assumptions.

In the future, we will also use the findings to model the evolution of public opinions and peer groups.

ACKNOWLEDGMENTS

Chen Lin is supported by the Natural Science Foundation of China (no.61472335). Yanghua Xiao is supported by National Key R&D Program of China under No.2017YFC0803700 and No.2017YFC1201200, by the National NSFC (No.61732004, No.61472085, No.U1509213, No.U1636207), by Shanghai Municipal Science and Technology project (No.16511102102, No.16JC1420401), Shanghai STCSMs R&D Program under Grant (No.16JC1420400). Chen Lin is the corresponding author.

REFERENCES

[1] B. M. Marlin and R. S. Zemel. Collaborative prediction and ranking with non-random missing data. In *RecSys* 2009, pages 5–12.
[2] H. Yang, G. Ling, Y. Su and MR. Lyu. Boosting response aware model-based collaborative filtering. *IEEE Transactions on Konwledge and Data Engineering*, 27(8):2064–2077, 2015.
[3] YD. Kim and S. Choi. Bayesian binomial mixture model for collaborative prediction with non-random missing data. In *RecSys* 2014, pages 201–208.
[4] JM. Hernández-Lobato, N. Houlsby and Z. Ghahramani. Probabilistic matrix factorization with non-random missing data. In *ICML* 2014, pages 1512–1520.
[5] P. Gopalan, JM. Hofman and DM. Blei. Scalable Recommendation with Hierarchical Poisson Factorization. In *UAI* 2015, pages 326–335.
[6] D. Liang, L. Charlin, J. Mclnerney and DM. Blei. Modeling user exposure in recommendation. In *WWW* 2016, pages 951–961.
[7] E. Noelle-Neumann. The spiral of silence a theory of public opinion. *Journal of communication*, 24(2):43–51, 1974.
[8] J. D. Kennamer. Self-serving bibias in perceiving the opinions of others. *Communication Research*, 17:393–404, 1990.
[9] J. Matthes. Observing the "spiral" in the spiral of silence. *International Journal of Public Opinion Research*, 27(2):155–176, 2014.
[10] RE. Anderson. Consumer dissatisfaction: The effect of disconfirmed expectancy on perceived product performance. *Journal of marketing research*, 38–44, 1973.
[11] W. De. Koster and D. Houtman. 'STORMFRONT IS LIKE A SECOND HOME TO ME' On virtual community formation by right-wing extremists. *Information, Communication & Society*, 11(8):1155–1176, 2008.
[12] E. Nekmat and WJ. Gonzenbach. Multiple opinion climates in online forums: Role of website source reference and within-forum opinion congruency. *Journalism & mass communication quarterly*, 90(4):736–756, 2013.
[13] P. Porten-Cheé and C. Eliders. Spiral of silence online: How online communication affects opinion climate perception and opinion expression regarding the climate change debate. *Studies in communication sciences*, 15(1):143–150, 2015.
[14] 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. *International Journal of Public Opinion Research*, 24(3):346–367, 2012.
[15] J. G. Carroll, F. H. Andrew, 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". *The Public Opinion Quarterly*, 61(3):452–463, 1997.
[16] HK. Meyer and B. Speakman. Quieting the Commenters: The Spiral of SilenceâĂŹs Persistent Effect on Online News Forums. *Quieting the Commenters: The Spiral of SilenceâĂŹs Persistent Effect*, 51, 2016.
[17] N. N. Dalvi, R. Kumar, and B. Pang. Para 'normal' activity: On the distribution of average ratings. In *ICWSM* 2013, pages 110–119.
[18] J. B. Cohen and M. E. Goldberg. The dissonance model in post-decision product evaluation. *Journal of Marketing Research*, 315–321, 1970.
[19] N. Jindal and B. Liu. Opinion spam and analysis. In *WSDM* 2008, pages 219–230.
[20] D. Godes and J. C. Silva. Sequential and temporal dynamics of online opinion. *Marketing Science*, 31(3):448–473, 2012.
[21] N. Hu, J. Zhang, and P. A. Pavlou. Overcoming the j-shaped distribution of product reviews. *Communications of the ACM*, 52(10):144–147, 2009.
[22] Liu, Y, Cao, X. and Yu, Y. Are You Influenced by Others When Rating?: Improve Rating Prediction by Conformity Modeling. In *RecSys* 2016, pages 269–272.
[23] P. Massa and P. Avesani. Trust-aware recommender systems. In *RecSys* 2007, pages 17–24.
[24] YL. Zhang, Q. Guo, J. Ni and JG. Liu. Memory effect of the online rating for movies. *Physica A: Statistical Mechanics and its Applications*, 417:261–166, 2015.
[25] J. McAuley, R. Pandey and J. Leskovec. Inferring networks of substitutable and complementary products. In *SIGKDD* 2015, pages 785–794.
[26] J. Tang, H. Gao and H. Liu. mTrust: discerning multi-faceted trust in a connected world. In *WSDM* 2012, pages 93–102.
[27] FM. Harper and JA. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19, 2016.
[28] P. McJones. Eachmovie collaborative filtering data set. *DEC Systems Research Center*, 249:57, 1997.
[29] P. DiMaggio, J. Evans and B. Bryson. Have American's social attitudes become more polarized?. *American journal of Sociology*, 102(3):690–755, 1996.
[30] RL. West and LH. Tumer. Introducing communication theory: Analysis and application. *McGraw-Hill Humanities/Social Sciences/Languages*, 2006.
[31] DA. Scheufle and P. Moy. Twenty-five years of the spiral of silence: A conceptual review and empirical outlook. *International journal of public opinion research*, 12(1):3–28, 2000.
[32] E. Neolle-Neumann. *The spiral of silence: Public opinion, our social skin.* University of Chicago Press., 1993.
[33] Zhang, Xiaoying and Zhao, Junzhou and Lui, John C.S. Modeling the Assimilation-Contrast Effects in Online Product Rating Systems: Debiasing and Recommendations *Proceedings of the Eleventh ACM Conference on Recommender Systems*
[34] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. *NIPS*, 20:1257–1264, 2008.
[35] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.