The Spiral of Silence and Its Application in Recommender System

Paper 118

ABSTRACT

Are people in recommender systems less willing to speak out if they perceive they are in the minority? We verify the famous "spiral of silence" theory in real recommender systems. We find that as the opinion climate gets stronger, the minority opinion holders are more incline to remain silent. Furthermore, we observe that (1) in the beginning, people have a tendency to hold back negative opinions and give positive responses. But the tendency is weakened for popular items, thus resulting in a spiral process where the average rating for each item firstly increases given more ratings received, slightly decreases after a certain time point, and finally converges. (2) People assess the "opinion climate" , by how people in their communities and the opinion leaders response. (3) A hardcore group of people, who are more active to reveal minority opinions, exists in every recommender system. Hardcore users are more prone to rate in extreme values. (4) People are more willing to demonstrate dissents for popular items and items that they are not interested on. People feel more obligated to praise a criticized item than the opposite, to criticize an appreciated item. The phenomenon of silent minority harms the performance of conventional recommender systems. We propose novel models which assume that the probability a user willing to express opinion is dependent on the rating and how the rating divergent to the perceived opinion climate. Furthermore, we utilize our empirical discoveries to guide the building of recommender models. We model the formation of community, opinion leaders, hardcore personality and item popularity to enhance recommendation performance. Experimental results demonstrate that our models outperform state-of-the-art recommendation models.

KEYWORDS

Spiral of Silence, Missing not at Random, Recommender System

1 INTRODUCTION

In 1974, German political scientist Elisabeth Noelle-Neumann proposed the Spiral of Silence Theory[1] [18]. The theory states that, due to the "fear of isolation", people are less willing to express their opinions if they perceive that they are not supported by the majority. It results in a spiral process in which the majority opinion receives growing popularity while other opinions are gradually pushed back. The process continues until the majority opinion ultimately becomes a social norm.

The theory has been acknowledged as "one of the most influential recent theories of public opinion formation" [10]. In the literature of mass communication and political science, many studies testify the theory. Typically, they conduct surveys and ask subjects to rate the willingness to speak out (i.e. enter a conversation, vote, donate, etc.) if their opinions are in the minority. However, this type of experiment protocols may be problematic, because the findings are based on hypothetical willingness instead of actual willingness [1]. The theory in its online form also triggers considerable critiques. Some researchers point out that, within the online context, factors such as anonymity might decrease the fear of isolation and thus empower "people in the minority to speak up more" [17].

Our first contribution in this work is to empirically verify the theory in terms of actual willingness to speak out in Recommender Systems (RS). We use several real data sets to analyze response patterns given how the users' opinion diverge from the perceived "opinion climate". Our study suggests that, especially in the beginning, negative feedback tend to self-censor, which triggers an upward spiral of average rating for each item.

We then further examine some key components in the theory. The key components include: (1) The perceptron of "opinion climate". The theory asserts that people use their "quasi-statistical" sense to assess current majority opinion. Our results reveal that, assessment of "opinion climate" might be related to the social group they identify themselves and the opinion leaders. (2) The existence of "hardcore groups". The theory presumes that some people are more active while the rest are more reluctant to respond. We empirically verify the existence of a consistent hardcore group. Hardcore users tend to choose more extreme rating values. (3) The strategies to remain silent. The theory presumes that users might choose different strategies to remain silent according to the nature of items. We observe that, item's popularity, personal interest and moral basis have significant impacts on the response patterns. Users are more prone to speak out for popular items, items that are not important to them, and to praise a criticized item than to criticize an appreciated item.

The phenomenon of "silent minority" will harm the performance of RS. Fig. 1 gives an illustrative example. Suppose user u_1 is a sensitive user who only responds to items on which he agrees with the majority. Given his responses, a conventional recommender will estimate his preferences as the average user preference. For example, a collaborative filtering recommender will determine u_1's nearest neighbor as u_2, u_3, as the similarities $s_{1,2}$ between u_1, u_2 and $s_{1,3}$ between u_1, u_3 are the highest. It leads to a predicted rating $r_{1,1} = \frac{s_{1,2}r_{2,1}+s_{1,3}r_{3,1}}{s_{1,2}+s_{1,3}} = 5$ for item v_1. As we can see from the example, the prediction is severely biased.

To address the challenge of "spiral of silence", one need to model the missing ratings as missing not at random observations(a.k.a. MNAR). In the RS community, MNAR models explicitly generate user response by user ratings [6, 22, 16]. Instead of optimizing

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

Conference'17, Washington, DC, USA
YYYY. 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

u/v	v_1	v_2	v_3	v_4	v_5
u_1	2	3		3	
u_2	5	3			2
u_3	5	3	5	3	
u_4			2	3	4
u_5	5		2		1

Figure 1: A toy example, ratings in gray are "hidden" opinions

$p(R|\Theta)$, which is the likelihood of ratings, MNAR models optimize $p(R, X|\Theta)$, where X is the set of *response* variables indicating whether a rating is missing. Consequently, MNAR models achieve better performance in predicting both ratings and responses.

In the nutshell, our models fall into the MNAR framework. However, for the best of our knowledge, we are the first to apply "spiral of silence" theory in recommender systems and consider the perception of support as a factor in an MNAR model. we model the possibility of user "speaking out" as dependent on users' perception of opinion climate. Our model adjusts the bias of response probability by connecting the rating values and opinion climate, thus improves the performances of traditional MNAR models, which are solely based on rating values. In the experiments on real data sets, our model achieves best results ($NDCG$@5 0.79) compared with other state-of-the-art MNAR models(best $NDCG$@5 0.7).

We also utilize the findings in our empirical study to build variant models. We incorporate hidden community, hardcore persona and item specific factors into the model. The comparative performances of model variants support the discoveries about the impact of social identity, hardcore groups and silent strategy on self definition of minority. We believe our study sheds insights into bridging social science and computer engineering.

The paper is organized as follows. In Sec 2 we briefly introduce the related works. In Sec 3 we present our empirical study on several real data sets. In Sec 4 we propose several model variants, based on the findings in Sec. 3. In Sec. 5 we demonstrate and analyze comparative performances of the models. Finally, in Sec. 6 we conclude our contribution and look into the future work.

2 RELATED WORK

Recently, recommendation system has attracted a lot of research attentions. In the fruitful literature, matrix factorization [12] techniques have exhibited superior performances in rating predictions. Probabilistic matrix factorization [21] implement the idea of matrix factorization from a probabilistic generative perspective. Opinion leaders and social communities can be included in this framework. For example, a SVD style model [15] factorizes the rating matrix to most representative users. The social recommender [8] models how a user's rating is affected by his/her trusted friends in a community.

When data is not missing at random, the probability of generating both observations and missing ratings (thus the "whole" data set) is not proportional to the likelihood of observed ratings. Marline et.al presented a pioneer work [16] of Missing Not At Random (MNAR) models. It assumes that response is a binary variable, which is generated by a Bernoulli distribution associated with the rating value. Along this line, several successive research works introduce new mechanism to generate responses from ratings. A continous rating is allowed in [14] with a step function. Such a soft assignment model is further improved in [23] with a mini batch algorithm. [11] adopts an OR operation over per-item, per-user, per-rating-value parameters. The most complicated model is given by [6], which proposes seperate generative process for the complete ratings and the responses, and cover the rating matrix by the response matrix as a mask to form the observations.

Another line is to consider a semi-observed variable "exposure". The response is determined by whether the item is exposed to the user, as well as the potential rating. A few recent works [13, 5] fall into this category. Instead of directly producing a response based on the hidden rating, an alternative approach is to probabilistically relate responses and ratings. For example, [19] presumes a parameter is involved in both processes of generating ratings and responses. Under the MNAR assumptions, conventional evaluation metrics, such as NDCG may fit better to response bias [20]. An approximate evaluation metric in top N observations is presented in [22].

The dynamics of public opinion is a long and refreshing research topic. Many empirical studies are performed in various domains. Some researchers have observed the trend of increasing average ratings. They offered several explanations. The first one is selection bias: users select and rate entities they are likely to like [3]. The second hypothesis is choice-supportive bias [2]: since users take too much efforts in finding a product, they will refuse to admit their poor judgment. The third hypothesis is that more reviews bring in more self-promotion spams [9]. On the contrary, some researchers find that later ratings are on average lower than earlier ratings [4]. The possible explanation is that, the volume of reviews restrict one's ability to diagnose previous reviews. Therefore when previous reviewers are very different, more reviews may lead to more purchase errors and lower ratings. Hu et.al observed a J-shaped distribution [7], presumably driven by purchasing bias (selecting higher product valuations) and under-reporting bias (report only when it is to "brag or moan").

Our empirical study shows that, pure positive ratings (3, 4 within a range of $[0, 5]$) are not dominant in recommender systems. The trend of average ratings is not globally monotone. We should also point out that, although the above biases make senses, we believe that our hypothesis in this paper is more reasonable. By introducing the "opinion climate", we provide a baseline for judging positive and negative opinions. Our experimental results prove that such a baseline improves models which consider the polarity of opinions simply by the rating values.

3 EMPIRICAL STUDY

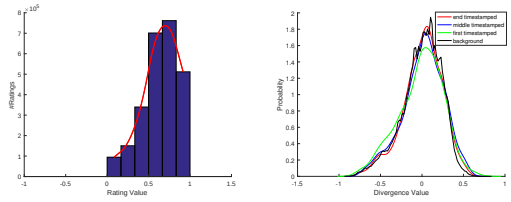
We use several real data sets, as shown in Tab. 1. The first one is Movielens1M, which is a collection of over 1M ratings on movies in the movielens website during a period of 34 months. The ratings are made on a 5-star scale. More than half of the ratings are positive feedbacks (4, 5). The second one is Eachmovie, which include nearly 2.6M ratings on films and videosis over a course of 18-months' experiment. The ratings are in the range of $(0, 1)$. Almost half of them are positive feedbacks (> 0.6). The two data sets are the most commonly used benchmarks in recommender systems.The ratings come with timestamps and tags.

We also use the recent Yahoo! data set, which is a set of ratings on songs through the Yahoo! web-scope data sharing program. The data set contains two subsets of ratings. The first (Yahoo!user) set consists of ratings supplied by users during normal interaction with Yahoo! Music services. The second source consists of ratings collected during an online survey, when each of the first 5400 users in Yahoo!user set was asked to provide exactly ten ratings for randomly selected songs. The ratings in Yahoo! dataset do not come with timestamps or tags.

The Yahoo! data sets provide unique opportunities to testify the spiral of silence theory. The random setting corresponds to a scenario where users are forced to response, against his actual willings. The user selected setting corresponds to a scenario where users are free to hide their responses. Therefore we can compare user's behaviors within different restrictions to study the key components of the theory.

3.1 Silent Minority

We use the common RS data sets to investigate the silent minority. We first show the historgram of ratings in each data set in Fig. 2(a) and Fig. 2(c) respectully. For each rating, we compute its difference to the current average ratings on the item. The rating divergence is defined as $d = r_{i,j,t} - r_{\hat{j},t}$, where $r_{i,j,t}$ is the rating by user i on item j at timestamp t, and $r_{\hat{j},t}$ is the average rating on item j by timestamp t. We report the distribution of d in three different time segments in Fig. 2(b) and Fig. 2(d). The *first*(green curve) distribution is based on the 95th to 100th ratings received for each item. The *end*(red curve) distribution is based on the last 5 ratings received for each item. For the *middle* distribution, we chronologically divide the ratings for each item into two parts, and the computation is excecuted on the last 5 ratings on the first devision. We also report a "background" distribution of $\bar{d} = r_{i,j,t} - \hat{r_j}$, where $\hat{r_j}$ is the average rating on item j at all times.

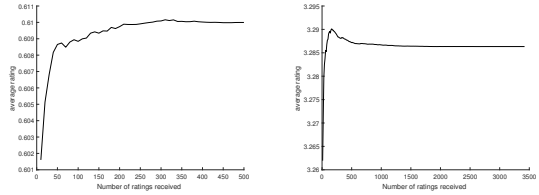


(a) r in Eachmovie (b) d in Eachmovie

(c) r in Movielens (d) d in Movielens

Figure 2: Distributions of rating value r and rating divergence d

We have the following observations in both the data sets. (1) The majority ($> 75\%$) of rating divergence falls in the small range of $[-1, 1]$ (in Eachmovie $[-0.2, 0.2]$). Minority opinion holders have a strong tendency to keep silent. (2) All distributions are left skewed. The mass of distribution of rating divergence locates at the right to the origin. (3) The timestamped distributions offer us a better opportunity to analyze the response patterns, as they are smoother and single peaked. We can see that as time goes by, the curves are becoming shallower and moving to the left. This suggests that, when the opinion climate is stronger (with more supporters), minority opinion holders are less likely to speak out. It also suggests that at the beginning, people are less likely to give negative feedback and more likely to give positive feedback. We will look into this issue in the next subsections.



(a) Eachmovie (b) Movielens

Figure 3: The spiral of average rating v.s. the number of ratings received

When users refuse to speak out relatively negative opinions, more positive feedback is observed while negative feedback is infrequently voiced. In theory this will lead to enlarging average rating. We plot the average rating for each item when they have $x = 10 \times k$ ratings, $k = 1, \cdots, 50$ for Eachmovie and $k = 1, \cdots, 350$ for Movielens. As shown in Fig. 3, two discoveries are clear. (1) An upward spiral is indeed activated. With more ratings received, the average rating for each item increases. (2) In the tail (approximately more than 350 ratings in Eachmovie and 500 in Movielens), the average rating begins to decrease. This trend suggests that for popular items, people are more incline to give negative feedback.

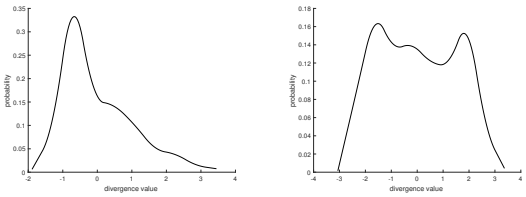
We have shown that the silent minority, defined as the group of users who hold different ratings with $d > \frac{\max(r)-\min(r)}{5}$, is universally observed in many recommender systems. For the yahoo! data set, there is no information about when the ratings are recorded. However, from the comparison between the random subset and user selected subset in Fig. 4, we see that, when users are not forced to express their opinions, the distribution is much wider, with multiple peaks at a large scale of $[-2, 2]$. Furthermore, they produce significantly less negative feedback and much more positive feedback. Thus our discovery of spiral of silent negative feedback in the common recommender systems is verified.

3.2 Opinion Climate

Opinion climate is a coined term to describe the mainstream opinion. For a thorough understanding of how the opinion climate develops and influences response patterns, we conduct the following studies on the Yahoo! data set.

Table 1: Statistics of the data sets

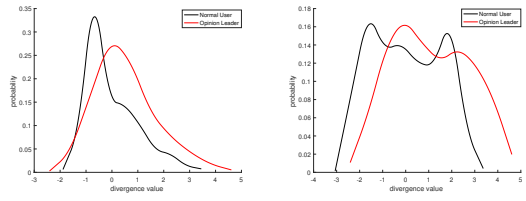
Dataset	#users	#Items	#Ratings	Positive	Negative	Time	#Tags
Movielens	6,040	3706	1,000,209	57.52%	42.48%	34 months	18
Eachmovie	61,131	1,622	2,559,107	54.72%	45.28%	18 months	10
Yahoo!user	15,400	1000	311,704	40.13%	59.87%	N/A	N/A
Yahoo!random	5400	1000	54,000	8.79%	91.21%	N/A	N/A



(a) Yahoo!random (b) Yahoo!user

Figure 4: The comparative probability distribution of rating divergence in Yahoo!

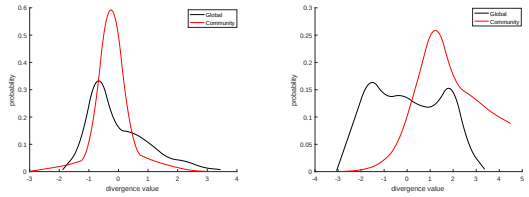
In the theory, opinion climate is highly affected by the mass media. In recommender systems, opinion leaders play a similar role as mass media. We select 10 most active users O (with largest number of ratings) as opinion leaders, and compute the average rating within opinion leaders $\hat{r_j} = \frac{\Sigma_{i \in O} r_{i,j}}{|O|}$. From Fig. 5, we see that, in the random setting, rating divergence to opinion leaders distribute as a normal distribution, while in the user selected setting, distribution of rating divergence to opinion leaders move to the right. The difference between two settings is more obvious than divergence to average people. It shows that opinion leaders magnify users' willingness to remain silence when they hold negative feedback and speak out positive opinions.



(a) Yahoo!random (b) Yahoo!user

Figure 5: The comparative distribution of rating divergence to opinion leaders in Yahoo!

On the Internet, with the infinite flow of information, users tend to have a limited vision of how others think and behave. They might focus on the opinions alike theirs, and thus perceive a so-called "looking-glass perceptron" of opinion climate. To study whether the opinion climate is dependent on user communities, we first represent each user as a rating vector on the item universe, and proceed to form a cluster c for users by selecting the KNN ($K = 50$) users with similar tastes. We then compute the average of community ratings $\hat{r_{j,c}} = \frac{\Sigma_{i \in c} r_{i,j}}{|c|}$ for community c. For each user u_i in community c, the rating divergence is $d = r_{i,j} - \hat{r_{j,c}}$.



(a) Yahoo!random (b) Yahoo!user

Figure 6: The comparative distribution of community specific rating divergence in Yahoo!

As shown in Fig. 6, we have the following observations. (1) In the random setting, rating divergence to community specific majority rating follows a shallow normal distribution, which centers at around 0. (2) In the user selected setting, distribution of rating divergence to community specific majority rating is single peaked and asymmetric. The probability peak is at around 1.25. The probability of negative feedback is much lower than rating divergence to global average rating. These observations strongly indicate that users perceive community specific opinion climate from his/her alike people.

3.3 Hardcore Users

In the theory, hardcore group is a bunch of people who are brave and willing to express different opinions. We define a "hardcore" score for each user, which is $h = \frac{n_i^h}{n_i}$, 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. In Eachmovie, high divergent ratings are $\{|r_{i,j} - \hat{r_j}| \geq 0.3\}$. In Movielens and Yahoo!users, $\{|r_{i,j} - \hat{r_j}| \geq 1.5\}$. In Yahoo!random $\{|r_{i,j} - \hat{r_j}| \geq 1.0\}$ as the random data set consists less ratings. We observe a small amount of hardcore users in all data sets. As shown in Tab. 2, the size of group reasonably decreases as the threshold of hardcore score increases.

Table 2: Percentage of hardcore groups

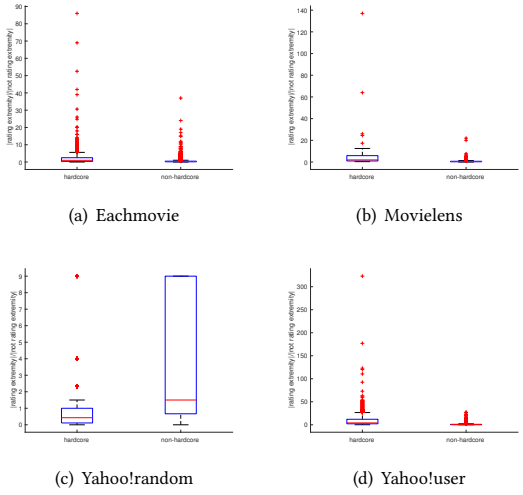
Dataset	$h = 0.5$	$h = 0.6$	$h = 0.7$	$h = 0.8$
Movielens	0.006	0.0026	0.00099	0.0005
Eachmovie	0.0674	0.03174	0.01693	0.01219
Yahoo!random	0.1672	0.0885	0.0470	0.0230
Yahoo!user	0.4234	0.2744	0.1668	0.0884

We first detect hardcore users in both yahoo data sets, and compare the hardcore users in two subsets. We report the overlap percentage $p_h = \frac{|G_h^{random} \bigcap G_h^{user}|}{|G_h^{random} \bigcup G_h^{user}|}$ in Tab. 3, where G_h^{random}, G_h^{user} are the set of hardcore users with hardcore score h in the random subset and the user selected subset respectfully. We also provide the overlap percentage in theory given that users are uniformly sampled by the probability distribution of hardcore shown in Tab. 2. We find that hardcore is a consistent personality of users. Despite of the threshold of hardcore score, the overlap percentage of hardcore users is significantly larger than the theoretical overlap, implying that users do not randomly decide to act hardcore.

Table 3: Percentage of hardcore group overlap

Threshold	$h \geq 0.5$	$h \geq 0.6$	$h \geq 0.7$	$h \geq 0.8$
Actural Overlap	0.16377	0.11737	0.0908	0.07194
Theretical Overalap	0.1169	0.0658	0.0341	0.0156

It is natural to relate hardcore with attitude certainty, while attitude certainty is represented by an extreme rating value. We then set $h \geq 0.5$ and plot the ratio of extreme ratings versus non-extreme ratings for hardcore and non-hardcore users in all data sets. Extreme ratings are $1, 5$ ratings in Movielens and Yahoo! and $0, 1$ ratings in Eachmovie. We can see from Fig. 7 that in three data sets, hardcore users have a higher median ratio of extreme ratings, and the tail is much longer. We also notice that when users are forced to express (the Yahoo!random set), the situation is reversed, hardcore users have a lower median ratio of extreme ratings and a short tail. This observation is again a supplementary evidence of the clear pattern that hardcore users are more likely to give extreme ratings.



(a) Eachmovie (b) Movielens

(c) Yahoo!random (d) Yahoo!user

Figure 7: Ratio of extreme ratings in hardcore and normal users

3.4 Hardcore and Items

Finally we study whether the strategies of holding back are dependent on the items. We have already seen the temporal changes in average ratings in common RS data sets in Fig. 3. In Movielens, the average rating first increases then decreases until converges. There is a similar trend in Eachmovie, although the downward incline is not steep in the tail. One possible reason is that some items are already popular before the Eachmovie project launches and starts to collect ratings. For a thorough study, we present in Fig. 8 the average ratings for popular items, which are items that receive more than 500 ratings. We see that for popular items, as the item receives more ratings, the average rating decreases.

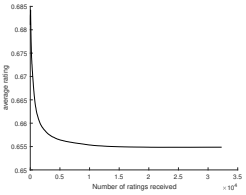
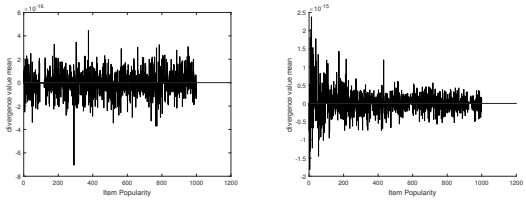


Figure 8: Average ratings of popular items in Eachmovie

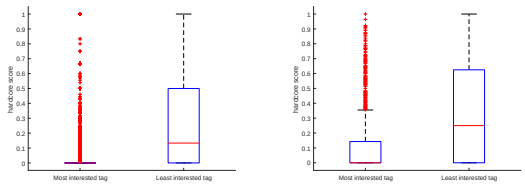
Next we study whether the responding is directly related to the popularity of items in Yahoo! data sets. We order the items in the Yahoo! data sets, according to the number of ratings assigned to them. A box plot is shown in Fig. 9 to reveal the distribution of rating divergence versus the order position of popularity. We can see that, in the random subset, the scales of rating divergence are almost the same for popular and unpopular items. In the user selected subset, the scale of rating divergence narrows as we choose less popular items. It shows that users adopt different strategies for popular items and unpopular items. When users are not restricted, they tend to speak out extreme opinions for popular items.



(a) Yahoo!random (b) Yahoo!user

Figure 9: The distribution of rating divergence upon item popularity

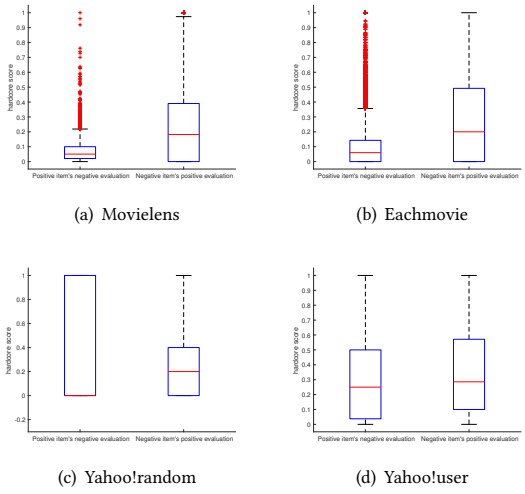
It is mentioned in the theory[18] that hardcore is related to personal interest or importance. To testify this assumption, we select the most rated tag and least rated tag (in number of ratings) for each user, and compute the hardcore score, where n_i is the number of ratings a user i gives to all items associated with the tag. As shown in Fig. 10, in both data sets, users are more willing to express different opinions (with a higher hardcore score) for least interesting items.



(a) Movielens (b) Eachmovie

Figure 10: The distribution of hardcore scores upon personal interest

Is hardcore 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 how people react in giving positive feedback ($r_{i,j} > \hat{r} + \frac{\max r - \min r}{p}$, $p = 5$ for Movielens, Eachmovie and Yahoo!users, $p = 10$ for Yahoo!random as this data set is much smaller) to items with average negative feedback ($\hat{r} < 3/5(\max r - \min r)$) and giving negative feedback ($r_{i,j} < \hat{r} - \frac{\max r - \min r}{n}$, $n = 5$ for Movielens, Eachmovie and Yahoo!users, $n = 10$ for Yahoo!random) to items with average positive feedback ($\hat{r} > 3/5(\max r - \min r)$) . As shown in Fig. 11, it is clear that people feel more obligated to save a criticized item than to underrate a highly appreciated item.



(a) Movielens (b) Eachmovie

(c) Yahoo!random (d) Yahoo!user

Figure 11: The distribution of hardcore scores under two moral situations

We also conduct analysis to correlate hardcore personality to user experience and context of ratings. No evidence is found to imply significant correlation between the probability of a user being hardcore and the number of ratings he/she gives, in which day of a week and which time of a day the rating is given, and the previous items he/she rates. We omit the analysis here due to the limit of space.

4 MODEL

As with most matrix factorization models, we assume there are K hidden aspects. The user preference is denoted as a vector $U_i \in R^K$, and each item's coefficient to the aspects is denoted as a vector $V_j \in R^K$. We assign a random variable $BU_i \in R$ to each user and $BV_j \in R$ to each item to represent the user bias and item bias. The rating given by user U_i to item V_j is denoted by $R_{i,j}$. The ratings are semi-observed. Whether the rating $R_{i,j}$ is observed is denoted by a binary response variable $X_{i,j}$, where $X_{i,j} = 1$ indicates the rating is observed and otherwise the rating is missing.

4.1 Model Base and COL

Intuitively, a user will give a high rating if the item matches his/her preference. If his/her opinion is close to the opinion climate, he/she is more likely to reveal this rating. Furthermore, our discovery in empirical study shows that users feel more obligated to give positive feedback on negative items. Taking into account of the above three intuitions, our **Base** model assumes the following three generative stages.

The preprocessing stage: For each user, generate the user preference from a Gaussian distribution with mean 0.

$$U_i \sim \mathcal{N}(0, \sigma_u^2) \tag{1}$$

Similarly, the user bias, item bias and item factors are also generated from Gaussian distributions. $BU_i, BV_j \sim \mathcal{N}(0, \sigma_b^2)$, $V_j \sim \mathcal{N}(0, \sigma_v^2)$.

The rating generation stage: The rating $R_{i,j}$ approaches to $U_i V_j + Bu_i + Bv_j$, with a zero-mean Gaussian error. Hence in this stage, generate the rating:

$$R_{i,j} \sim \mathcal{N}(U_i V_j + BU_i + BV_j, \sigma_r^2) \tag{2}$$

The response generation stage: $X_{i,j} = 1$ has a higher probability if the rating divergence to opinion climate $R_{i,j} - E_j$ is small. Generate the response value $X_{i,j}$ from Equ.3

$$P(X_{i,j} = 1 | R_{i,j}, E_j, \tau) = \frac{1}{1 + \exp\left(-\tau(R_{i,j} - E_j)\right)} \tag{3}$$

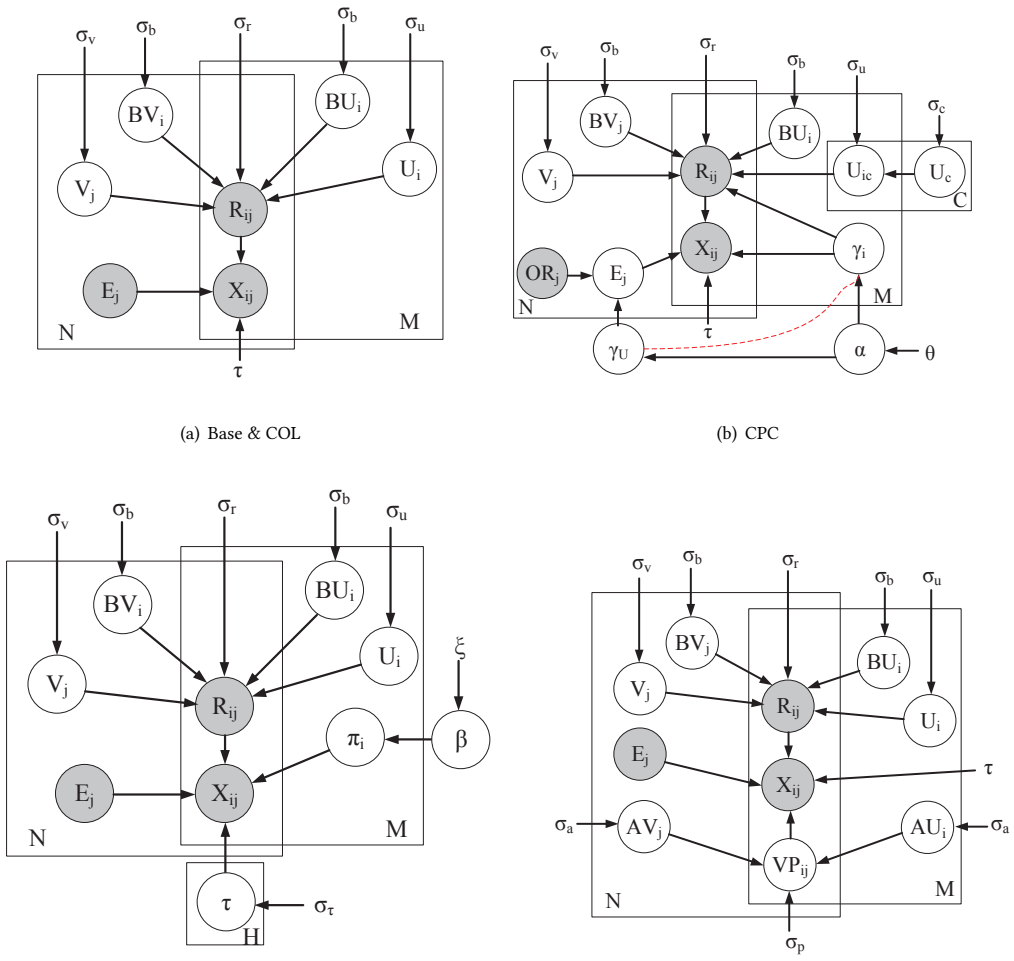
where τ is a hardcore strength parameter, E_j is the opinion climate for the particular item j. In the Base model, we set the opinion climate to be the average of all observed ratings.

$$E_j = \frac{\Sigma_j R_{i,j} X_{i,j}}{\Sigma_j X_{i,j}} \tag{4}$$

We next present four variant models, each individually models one important discovery in our empirical study. With a simple modification in the Base model, we obtain the **Conditional on Opinion Leader (COL)** model. The graphical representation of Base and COL is illustrated in Fig. 12(a). Instead of computing a global majority rating by Equ.4, the average is taken over some expert users e.

$$E_j = \frac{\Sigma_e R_{e,j} X_{e,j}}{\Sigma_e X_{e,j}} \tag{5}$$

where the experts e are extracted by some expert identification algorithms. This allows the flexibility of utilizing side information sources, such as cascaded social networks, to recognize opinion leaders.



(a) Base & COL

(b) CPC

(c) CPP

(d) CPV

Figure 12: Graphic representation of models

4.2 Model CPC

Our empirical discovery suggests a user perceives a filtered opinion climate of people sharing similar preferences. In the **Conditional Probability on Community (CPC)** model, we introduce a random $C-$dimensional vector γ_i to represent the community assignment of each user, $\gamma_{i,c} \in \{0,1\}, \Sigma_c \gamma_{i,c} = 1$.

Intuitively, users in the same community are alike. We assume that the "common" preference in community c, denoted by U_c. Thus as summarized in Fig. 12(b), the *preprocessing stage* is as follows.

Generate C communities. For each community, U_c is generated from a zero-mean Gaussian distribution $U_c \sim \mathcal{N}(0, \sigma_u^2)$. Generate a global community distribution $\alpha \sim Dir(\theta), \Sigma_c \alpha_c = 1, \forall \alpha_c > 0$. For each user, first generate the community indicator $\gamma_i \sim Discrete(\alpha)$, then generate the user preference according to his/her community:

$$U_i \sim \Pi_c \mathcal{N}(U_c, \sigma_u^2)^{\gamma_{i,c}} \tag{6}$$

Suppose OR_j represents the observed ratings, $OR_{i,j} = R_{i,j} X_{i,j}$, in the *response generation stage*, we replace the opinion climate in Equ.4 by a community specific opinion climate.

$$E_{c,j} = \frac{\Sigma_j \gamma_{i,c} OR_{i,j}}{\Sigma_i \gamma_{i,c} X_{i,j}} \tag{7}$$

4.3 Model CPP

To model the split of users between hardcore and normal groups, the **Conditional Probability on Persona (CPP)** model, shown in Fig. 12(c), introduces a persona variable, denoted by $\pi \in \{0,1\}$. Intuitively, when a user is hardcore $\pi_{i,0} = 1$, he/she is more likely to speak out regardless of the opinion climate. Thus the strength parameter τ is reasonably smaller. Model Base,COL, CPC and CPV use a global τ for every user. In model CPP, τ is personalized for each persona.

In the *preprocessing stage*, draw a persona distribution from a Beta distribution $\beta \sim \mathcal{B}(\xi), 0 < \beta < 1$. For each persona, Draw a hardcore coefficient $\tau_z \sim \mathcal{N}(1, \sigma_\tau), z \in \{0, 1\}$. For each user, draw a persona from a Bernouli distribution $\pi_i \sim Bern(\beta)$.

The response probability is dependent on the user's persona π_i, the hardcore strengths τ for each persona. In the *response generation stage*, generate the response value $X_{i,j}$ from Equ.8

$$P(X_{i,j} = 1|R_{i,j}, E_j, \pi_i, \tau) = \Pi_z \frac{1}{1 + \exp(-\tau_z(R_{i,j} - E_j))}^{\pi_{i,z}} \quad (8)$$

4.4 Model CPV

Finally we give model **CPV** to associate the silence strategy to the nature of items. Our empirical findings reveal that users are less hardcore when it comes to his/her favorite items and niche items. Intuitively, in the generation of response, there is a hidden variable that interacts with both users and items. As shown in Fig. 12(d), in the *preprocessing stage*, we introduce auxiliary variable $AU_i \in R^K$ to represent user attention on K categories, $AV_j \in R^K$ to represent item's "absence". We have $AU_i, AV_j \sim \mathcal{N}(0, \sigma_a^2)$.

In the *rating generation stage*, we have $VP_{i,j}$ to represent item-user-specific popularity. in addition to Eq. 2, let $VP_{i,j}$ to be sampled by:

$$VP_{i,j} \sim \mathcal{N}(AU_i AV_j, \sigma_p^2) \quad (9)$$

Intuitively, if a user pays more attention to a category, and the item is absent from the best seller in the category, then it is less likely a different rating $r_{i,j}$ to be voiced. Consequently, we modified the response probability to

$$P(x_{i,j}|r_{i,j}, E_j, \tau) = \frac{1}{1 + \exp(-\tau(r_{i,j} - E_j - VP_{i,j}))} \quad (10)$$

4.5 Inference

The update equations in the four models share a similar form. For computational convenience, we first obtain $B_{i,j,c}$ and $p_{i,j,c}$ for each user, item and community at an iteration. Then we update U_i and V_j as follows.

$$\begin{aligned}
&p_{i,j,c} = U_{i,c}^T V_j + BU_i + BV_j \\
&B_{i,j,c} = (R_{i,j} - p_{i,j,c})^{[X_{i,j}=1]}(-\tau \mathcal{LS}(-\tau(p_{i,j,c} - E_{c,j})))^{[X_{i,j}=0]} \\
&U_{i,c} \leftarrow U_{i,c} + lr(\hat{\gamma}_{i,c} \sum_{j=1}^{N} B_{i,j,c} V_j - (U_{i,c} - U_c)) \\
&V_j \leftarrow V_j + lr(\sum_{c=1}^{C} \sum_{i=1}^{M} \hat{\gamma}_{i,c} B_{i,j,c} U_{i,c} - V_j)
\end{aligned} \quad (11)$$

where $\mathcal{LS}(x) = \frac{1}{1+\exp(-x)}$ is a logistic function, lr is the learning rate. For model Base and COL, the c subscript is ignored (i.e. $p_{i,j,c}$ to $p_{i,j}$, $B_{i,j,c}$ to $B_{i,j}$), and set $\hat{\gamma}_{i,c} = 1$.

For model CPC and CPP, we apply EM algorithms to infer the model parameters. . in the E-step we need to compute $\omega_{i,j,*}$ for each user, item and community c ($* = z$ for persona in CPP).

$$\omega_{i,j,*} = \begin{cases} P(X_{i,j} = 1|R_{i,j}, E_j)P(R_{i,j}) & X_{i,j} = 1 \\ 1 - P(X_{i,j} = 1|R_{i,j}, E_j) & X_{i,j} = 0 \end{cases}$$

The hidden community c is then inferred by

$$\hat{\gamma}_{i,c} \leftarrow \frac{\alpha_c \prod_{j=1}^{N} \omega_{i,j,c}}{\sum_{c=1}^{C} \alpha_c \prod_{j=1}^{N} \omega_{i,j,c}} \quad (12)$$

For model CPP, the update equations are given by:

$$\begin{aligned}
&\hat{\pi}_{i,z} \leftarrow \frac{\beta_z \prod_{j=1}^{N} \omega_{i,j,z}}{\sum_{z=0}^{1} \beta_z \prod_{j=1}^{N} \omega_{i,j,z}} \\
&B_{i,j,z} = (R_{i,j} - p_{i,j})^{[X_{i,j}=1]}(-\tau_{i,z} \mathcal{LS}(-\tau_{i,z}(p_{i,j} - E_j)))^{[X_{i,j}=0]} \\
&U_i \leftarrow U_i + lr(\sum_{z=0}^{1} \hat{\pi}_{i,z} \sum_{j=1}^{N} B_{i,j,z} V_j - U_i) \\
&V_j \leftarrow V_j + lr(\sum_{z=0}^{1} \sum_{i=1}^{M} \hat{\pi}_{i,z} B_{i,j,z} U_i - V_j)
\end{aligned} \quad (13)$$

For model CPV, we ignore the c subscript, set $\hat{\gamma}_{i,c} = 1$ and replace $E_{c,j}$ by $E_{c,j} = E_j + VP_{i,j}$, where $VP_{i,j}$ is updated by

$$\begin{aligned}
&D_{i,j} = \tau(\mathcal{LS}(-\tau(R_{i,j} - E_j)) - 1)^{[X_{i,j}=1]}(\mathcal{LS}(-\tau(p_{i,j} - E_j)))^{[X_{i,j}=0]} \\
&AU_i \leftarrow AU_i + lr(\sum_{j=1}^{N} D_{i,j} AV_j - AU_i) \\
&AV_j \leftarrow AV_j + lr(\sum_{i=1}^{M} D_{i,j} AU_i - AV_j) \\
&VP_{i,j} \leftarrow AU_i^T AV_j
\end{aligned} \quad (14)$$

5 EXPERIMENT

The data sets used in this section include Yahoo!, movieLen-1M, and eachmovie mentioned above. In addition, we use the Mtweet benchmark with 106, 337 ratings by 3, 972 users on 2, 043 movies crawled from Social Network. In each data set, we remove users whose number of ratings is less than 10, and items which have less than 10 ratings. For each user, we randomly select 5 ratings to construct the test data sets. All ratings are converted to the range of $(1, 5)$.

5.1 Comparative Study

The major evaluation metric is *NDCG@L*, which is a standard measure for ranking systems. NDCG@L is short for the normalized discounted cumulative gain for top L results.

$$NDCG@L = \sum_{i=1}^{M} \frac{\sum_{l=1}^{L} (2^{r_i(p(l,i))} - 1)/log(1 + l)}{M \sum_{l=1}^{L} (2^{r_i(t(l,i))} - 1)/log(1 + l)} \quad (15)$$

where $p(l, i)$ is the index of the test items sorted in descending order by predicted ratings, $t(l, i)$ is the index of the test items sorted in descending order by true ratings.

The comparative *NDCG* is conducted on Yahoo!random data set. Evaluating *NDCG@L* on a non-randomly missing data set, such as Yahoo!random, has been used as the primary criteria in many MNAR researches [6, 16].

In all our models, we set the aspect numbers $K = 5$, community numbers $C = 2$, hyper-parameters $\tau = 1$ for model Base,COL,CPC and CPV. $\xi = \theta = 2$, All the variances are set to $\sigma_b = \sigma_a = \sigma_r = \sigma_u = \sigma_v = \sigma_\tau = 1$. In COL, the top 10 users with most ratings on each item are selected as opinion leaders for this item. The learning rate is set to $lr = 0.0001$ in Yahoo!, Mtweet and movieLen-1M, $lr = 0.001$ in Eachmovie. Convergence is obtained when the change in mean log-likelihood $\Delta\bar{L} < 0.0001$ or stopped at maximal 1500 rounds.

We compare our models to a wide range of available models, including conventional memory-based and model-based collaborative filtering recommenders, MNAR models, and ranking based models. The comparative models include (1)uKNN: the user based K-Nearest Neighbor collaborative filtering recommender; (2) MF: the standard matrix factorization model [12]; (3)PMF: the probabilistic matrix factorization model [21]; (4)BPR: a classic ranking based model [24] (4)CPT-v and Logit-vd: the first MNAR models [16] with $K = 10$; (5) MF-MNAR [6]: the probabilistic MNAR model which masks the rating matrix by a response matrix, we set $K = 20$; (6)RAPMF [23]: an MNAR model which incorporates users' response models into the probabilistic matrix factorization (parameter $K = 5$). The variances for the above models are tuned by cross validation.

As shown in Fig. 13(a) and Fig. 13(b), we can see that (1) The four variants of our model outperform all non ranking-based models in terms of NDCG at different lengths, which demonstrates the dominant advantages of adjusting a opinion climate in the MNAR models. (2) CPC performs consistently best in all NDCGs, even better than ranking based models, proving the compactly of our approach. (3) COL, CPP and CPV are better than the base model, which verifies our empirical findings. (4) COL, CPP and CPV perform slightly worse than CPC. A possible explanation is that the user's perceives of opinion climate is mostly affected by the community. (5) BPR is the second best in all NDCGs. Because BPR directly optimizes the ranking of items, it is reasonable that it outperforms other rating optimization models.

5.2 Parameter Tuning

We have three important parameters in the CPC models τ, K and C. In order to see the effects of these parameters, we set $\tau = 0.2, 0., 5, 1.0, 1.5, 2.0$, $K = 5, 10, 15, 20$ and $C = 1, 2, 3, 4$. The *NDCG*@5 result of model CPC is shown in Fig. 14. We can conclude that (1) the performance is quite stable. No matter how we set the parameters, $NDCG@5 > 0.74$, which is better than all non-ranking based recommender systems. This again demonstrate the superiority of our models. (2) The best result is achieved when $\tau = 1.5$. It suggests that the hardcore effect is relative strong as $\tau > 1$. (3) *NDCG*@5 results are high for small values of K, C. However, when K and C are too large, the performance is harmed due to over-fitting.

5.3 Perceived Opinion Climate

In all models, the perceived opinion climate is based on existing ratings. To give a quantitive sense of the necessity to build opinion climate upon observations, we modify Equ. 7 as $E_{c,j} = U_c V_j$. In this way, the opinion climate is based on mean user preference within the community. As shown in Fig. 15, such a modification leads to worse *NDCG@L* performance at every length L of the recommendation list. It shows that, users need to compare their own opinions to the opinion climate they observed, instead of to the opinion climate they inferred.

5.4 AUC Performance

Finally, we provide a supplementary experiment on the AUC performance. AUC is the area under recall-precision curve, which is a common measure to evaluate the performance of a binary classifier. We should note that AUC does not reflect how the RS captures MNAR ratings. Therfore, AUC is not the primary measurement for MNAR models. However, as shown in Tab. 4, our models achive best results in two data sets (Yahoo! and Mtweet) and second best results in the remaining two data sets (Movielens and Eachmovie). Thus our models are comparable to the state-of-the art models in term of AUC evaluation. Furthermore, we observe that CPC and COL are comparable when evaluated by AUC. The underlying reason is that, COL may not be as good as CPC in predicting the top results. But in the long run, the opinion leaders have strong influence over user behaviors.

6 CONCLUSION

In this contribution we verify the "spiral of silence" theory in real recommender systems. We use the empirical discovies to build recommender models that outperform state-of-the-art recommenders. In the future, we plan to investigate how the empirical discoveries can be combined to improve performances.

REFERENCES

[1] 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.

[2] J. B. Cohen and M. E. Goldberg. The dissonance model in post-decision product evaluation. *Journal of Marketing Research*, pages 315–321, 1970.

[3] N. N. Dalvi, R. Kumar, and B. Pang. Para 'normal' activity: On the distribution of average ratings. In *(ICWSM)*, pages 110–119, 2013.

[4] D. Godes and J. C. Silva. Sequential and temporal dynamics of online opinion. *Marketing Science*, 31(3):448–473, 2012.

[5] P. Gopalan, J. Hofman, and D. Blei. Scalable recommendation with hierarchical poisson factorization. In *UAI*, 2015.

[6] J. M. Hernandez-Lobato, N. Houlsby, and Z. Ghahramani. Probabilistic matrix factorization with non-random missing data. In *ICML*, 2014.

[7] N. Hu, J. Zhang, and P. A. Pavlou. Overcoming the j-shaped distribution of product reviews. *Commun. ACM*, 52(10):144–147, Oct. 2009.

[8] M. Jamali, T. Huang, and M. Ester. A generalized stochastic block model for recommendation in social rating networks. In *RecSys* 2011, pages 53–60.

[9] N. Jindal and B. Liu. Opinion spam and analysis. In *WSDM* 2008, pages 219–230.

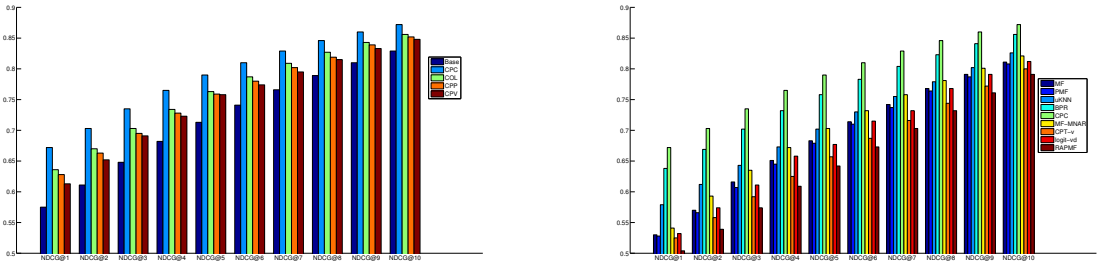
[10] J. D. Kennamer. Self-serving bibias in perceiving the opinions of others. *Communication Research*, 17:393–404, 1990.

[11] Y.-D. Kim and S. Choi. Bayesian binomial mixture model for collaborative prediction with non-random missing data. In *RecSys* 2014, pages 201–208.

[12] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

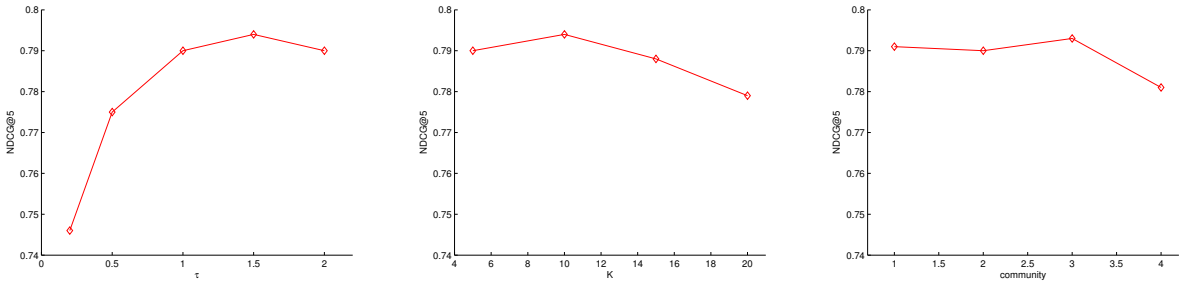
[13] D. Liang, L. Charlin, J. McInerney, and D. M. Blei. Modeling user exposure in recommendation. In *WWW* 2016, pages 951–961.

[14] G. Ling, H. Yang, M. R. Lyu, and I. King. Response aware model-based collaborative filtering. In *UAI*, pages 501–510, 2012.



(a) Our models (b) CPC v.s. state-of-the-art models

Figure 13: comparable experiment NDCG performance at top L items



(a) τ (b) K (c) C

Figure 14: NDCG@5 performance of CPC over different values of parameters

Table 4: Comparative AUC Performances

	Base	CPC	COL	CPP	CPV	uKNN	PMF	CPT-v	logit-vd	MF-MNAR	RAPMF
Yahoo	0.803	**0.880**	**0.882**	0.857	0.345	0.776	0.504	0.825	0.841	0.863	0.535
Mtweet	0.823	**0.864**	**0.863**	0.832	0.823	0.766	0.556	0.828	0.832	0.856	0.582
MovieLen	0.874	0.898	**0.902**	0.871	0.517	0.818	0.705	0.878	0.882	**0.929**	0.662
Eachmovie	0.840	**0.905**	0.900	0.864	0.776	0.739	0.603	0.792	0.821	**0.912**	0.562

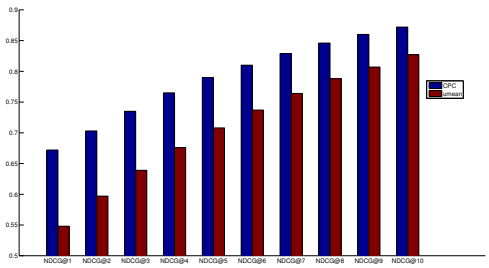


Figure 15: Performance of CPC on perceived opinion climate and inferred opinion climate

[15] N. N. Liu, X. Meng, C. Liu, and Q. Yang. Wisdom of the better few: Cold start recommendation via representative based rating elicitation. In *RecSys* 2011, pages 37–44.

[16] B. M. Marlin and R. S. Zemel. Collaborative prediction and ranking with non-random missing data. In *RecSys* 2009, pages 5–12.

[17] M. McDevitt, S. Kiousis, and K. Wahl-Jorgensen. Spiral of moderation: Opinion expression in computer-mediated discussion. *International Journal of Public Opinion Research*, 15(4):454–470, 2003.

[18] E. Neolle-Neumann. *The spiral of silence: Public opinion, our social skin*. University of Chicago Press., 1993.

[19] S. Ohsawa, Y. Obara, and T. Osogami. Gated probabilistic matrix factorization: Learning usersfk?attention from missing values. In *IJCAI*, pages 1888–1894, 2016.

[20] B. Pradel, N. Usunier, and P. Gallinari. Ranking with non-random missing ratings: Influence of popularity and positivity on evaluation metrics. In *RecSys* 2012, pages 147–154.

[21] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. *NIPS*, 20:1257–1264, 2008.

[22] H. Steck. Training and testing of recommender systems on data missing not at random. In *SIGKDD* 2010, pages 713–722.

[23] H. Yang, G. Ling, Y. Su, M. R. Lyu, and I. King. Boosting response aware model-based collaborative filtering. *IEEE Transactions on Knowledge and Data Engineering*, 27(8):2064–2077, Aug 2015.

[24] Rendle, S.; Freudenthaler, C.; Gantner, Z. and Schmidt-Thieme, L. BPR: Bayesian Personalized Ranking from Implicit Feedback In *UAI*, 2009, 452-461