Social Influence Bias in Recommender Systems: A Methodology for Learning, Analyzing, and Mitigating Bias in Ratings

Sanjay Krishnan, Jay Patel, Michael J. Franklin, Ken Goldberg
Department of Electrical Engineering and Computer Sciences, UC Berkeley
{sanjaykrishnan, patel.jay, franklin, goldberg}@berkeley.edu

ABSTRACT

To facilitate browsing and selection, almost all recommender systems display an aggregate statistic (the average/mean or median rating value) for each item. This value has potential to influence a participant's individual rating for an item due to what is known in the survey and psychology literature as Social Influence Bias; the tendency for individuals to conform to what they perceive as the "norm" in a community. As a result, ratings can be closer to the average and less diverse than they would be otherwise. We propose a methodology to 1) learn, 2) analyze, and 3) mitigate the effect of social influence bias in recommender systems. In the Learning phase, a baseline dataset is established with an initial set of participants by allowing them to rate items twice: before seeing the median rating, and again after seeing it. In the Analysis phase, a new non-parametric significance test based on the Wilcoxon statistic can quantify the extent of social influence bias in this data. If this bias is significant, we propose a Mitigation phase where mathematical models are constructed from this data using polynomial regression and the Bayesian Information Criterion (BIC) and then inverted to produce a filter that can reduce the effect of social influence bias. As a case study, we apply this methodology to the California Report Card (CRC), a new recommender system that encourages political engagement. After the Learning phase collected 9390 ratings, the non-parametric test in the Analysis phase rejected the null hypothesis, identifying significant social influence bias: ratings after display of the median were on average 19.3% closer to the median value. In the Mitigating phase, the learned polynomial models were able to predict changed ratings with a normalized RMSE of 12.8% and reduce bias by 76.3%. Results suggest that social influence bias can be significant in recommender systems and that this bias can be substantially reduced with machine learning. The CRC, our data, and experimental code can be found at:
http://californiareportcard.org/data/

1. INTRODUCTION

In the 1950's, Solomon Asch performed a well-known set of experiments [2,3,7] where participants were asked to choose which of a set of lines matched the length of reference lines. When working in private, only 1% of answers were incorrect. But when answering in the presence of a group of confederates who agreed on incorrect answers, 25% of participants conformed to the incorrect consensus values. These results were widely repeated to confirm what is now known as *social influence bias*: the answers of other participants encourage "conformity" – responses similar to the community "norm" [11,21,28].

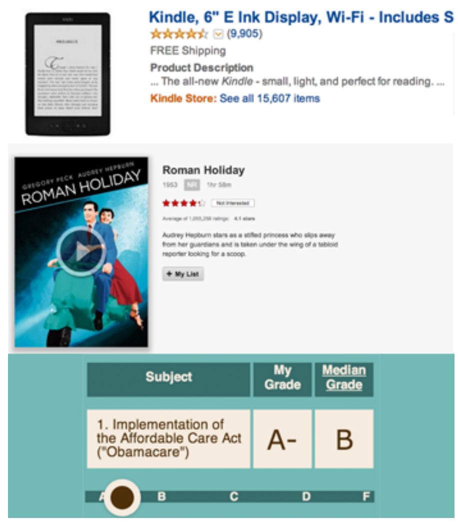

Figure 1: Typical displays of aggregate statistics (the average/mean or median rating value) in Amazon, Netflix, and the California Report Card that can lead to social influence bias. This paper explores a methodology for learning, analyzing, and mitigating such bias.

This effect can occur in almost all recommender systems, which display the "norm" aggregate statistics (the average/mean or median rating values) for items before asking participants to enter their own rating values, which is of course reasonable to facilitate browsing and selection. For example, online retailers such as Amazon display the average rating value for products and Netflix displays the average rating value of movies (Figure 1). Display of average ratings values can also be used as an incentive [19] to reveal information about peers after a participant enters his or her own grade. Display of such data also increases the perceived transparency of open democracy platforms that encourage political engagement [1,24,25].

Social influence bias can yield ratings that are closer to the average, less diverse, and less representative of participants' true evaluations for items, which can in turn produce bias in similarity measures between items and users. In this paper, we propose a methodology to learn, analyze, and mitigate the effects of social influence bias in recommender systems. As a case study, we apply our techniques to a new recommender system, the California Report Card (CRC), where participants assign letter grades (A+ to F, a 13 point scale) to the State of California on six political issues. The CRC interface reveals median grade values to participants after they enter their own ratings and then allows participants to revise their ratings.

The key insight is that the combination of initial and revised ratings pairs allows us to determine if the social influence bias is statistically significant, and if so, can be used to build an inference model that can mitigate the effects of social influence bias. The methodology includes three phases:

-Learn To initialize with baseline data, an initial "learning" phase asks an initial set of participants to rate a set of items twice: before seeing the median rating, and again after the median is revealed. This collects triplets of ratings for each participant (initial rating, median rating, and final rating).

-Analyze Given these triplets, we propose a new non-parametric significance test based on the Wilcoxon statistic to determine whether ratings that were changed are significantly closer to the median, i.e. the degree of social influence bias for each item.

-Mitigate Using the Bayesian Information Criterion (BIC), we learn a polynomial function of optimal degree that estimates the initial rating from the final rating and the median. This can be used in a post-learning phase (when medians are always visible), or on historical ratings, to estimate what a participant's rating would be without social influence bias.

In our case study, we apply this methodology to the California Report Card (CRC), a two-part recommender system that encourages political engagement. In Part I, participants assign letter grades (a 13-point rating scale) to the state of California on six political issues. Part I uses the six ratings to quantify similarity between participants and issues. In Part II, participants enter textual suggestions about new political issues and grade the suggestions of other participants. Part II uses participant ratings to identify (recommend) valuable suggestions. The CRC was announced via press and social media in late January 2014.

Results to date from the CRC suggest that given the opportunity, many participants will revise their grades/ratings: 862 out of 9390 ratings were changed after participants saw the median value. We found statistically significant effects of social influence bias, with ratings on average 19.3% closer to the median value than ratings that were not changed. We also conducted an independent reference survey using SurveyMonkey to ask a random sample of 611 participants from the company's paid pool of California participants to grade the same set of issues without displaying the median values. This data did not exhibit the same clustering around the median as the CRC, which comparably had ratings that were statistically significantly closer to the median (12.0%), suggesting that social influence bias is an important factor.

As described in the next section, earlier studies of social influence bias in recommender systems have focused on binary ratings (eg. up or down) [22,29]. Since many recommender systems have multi-valued rating scales (eg. 5 stars), we explore the effect on multi-valued ratings and develop a non-parametric significance test that avoids assumptions about the distribution of ratings. We then show how machine learning can be used to estimate unbiased ratings and present results with data from the California Report Card, including learning curves that show that in most cases estimation converges relatively quickly. Applying the proposed methodology to existing recommender systems raises a number of interesting questions for further research.

2. RELATED WORK

The Asch model for conformity is the theoretical basis for what is sometimes called *social herding*, the tendency to conform [4,5], and this has been a popular consumer choice model in economics [9,12,16]. Such models have also been studied in psychology as "persuasion bias" [11]. In 2011, Lorenz et al. described how these biases can undermine the effectiveness of crowd intelligence in estimation tasks [18]. They argue that movement towards the group consensus causes a diminished diversity of opinion potentially leading to inefficiencies and inaccurate collective estimates. Danescu-Niculescu-Mizil et al. analyze helpfulness ratings on Amazon product reviews [10]. They found that the helpfulness ratings did not just depend on the content of the review but also its aggregate score and its relationship to other scores. In order to better distinguish social influence from other biases, Muchnik et. al designed a randomized experiment in which comments in an online forum were randomly up-treated or down-treated [22]. They concluded a statistically significant bias where a positive treatment increased the likelihood of positive ratings by 32%. In both Danescu-Niculescu-Mizil et al. and Muchnik et al., they looked at the problem of Social Influence bias in an a priori setting, where users see the aggregate statistic before giving their rating. Our work tests for a particular form of social influence where users are given the opportunity to change their opinions following the feedback.

Zhu et al. conducted an experiment in which users evaluate an image on a subjective question with binary scale (eg. "Is this image cute?"), which was followed (either immediately or later) by a presentation of the crowd consensus opinion [29]. Users were given an opportunity to change their response, and they concluded that there was a significant tendency to change submissions. The tendency to change was the strongest when users were asked to make their second decision much later and not immediately after the first. However, Zhu et al. also acknowledge there are competing psychological factors at work in this experiment. Along these lines, Sipos et al. argue that context along with an aggregate rating plays a large role in the users' ratings. That is, users may attempt to "correct" the average, by voting in a more polarizing manner (more positively or negatively) [27]. We extend this prior work to measure and predict these changes when the input is more complex than a binary scale, and propose a non-parametric methodology that can be, in principle, extended to a variety of different input mechanisms. Our model can also account for a changing aggregate statistic such as a median rating changing as more data is collected.

3. LEARNING PHASE

In this section, we describe the learning phase of our technique where we collect the triplets (initial rating, final rating, and observed median) for building our model. We will explain in detail the system design of the California Report Card, how we record changed ratings, and define the notation that we will use in the following sections.

3.1 The California Report Card

The California Report Card (CRC) [1] is a prototype cross-platform web/mobile application designed to allow participants to advise California state leaders on timely policy issues. The CRC extends our earlier work with Opinion Space and Eigentaste [6,13–15,23]. In the CRC, participants assign letter grades (A+ to F) to the state of California on the following six issues: (1) Implementation of the Affordable Care Act ("Obamacare"), (2) Quality of K-12 public education, (3) Affordability of state colleges and universities, (4)

[1]This study was approved by our Human Subjects committee as per IRB Protocol 2014-01-5918.

Access to state services for undocumented immigrants, (5) Laws and regulations regarding recreational marijuana, and (6) Marriage rights for same-sex partners. Grades (Ratings) are assigned on a thirteen point scale (A+,A,A-,...,D-,F). These issues are posed in a fixed order each with the same input scale. Participants submit ratings using a click-and-drag slider interface as illustrated in Figure 2. On mobile devices, participants touch and drag to indicate the desired rating.

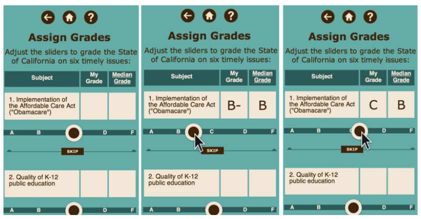

Figure 2: After entering their rating, the median rating over all participants is revealed. Participants have the option to change their rating after seeing the median.

Upon release of the slider, the CRC reveals the median for that issue over all prior participants. Even after the median is revealed the slider is still active and participants can change their ratings. However, it is important to note that participants were not explicitly told that they could change their rating. Another important observation is that participants who accessed the application at different times may have seen different medians as they were calculated based on the data up to that point. We recorded the initial rating, the median that the participant observed, and any subsequent changes along with timestamps for each of the events. Rating all of the six issues was not mandatory and participant had the option to skip any of the issues. To analyze this data, we mapped these 13 grade values linearly onto a scale from 0 to 1, with 1 being an A+ and 0 being an F.

3.2 Notation

Let P denote the set of all participants. For each participant $p_j \in P$, we associate a 3-tuple of ratings ($g_i[j]$, $m[j]$, $g_f[j]$) which represent the initial rating, median observed by the participant, and the final rating. For each issue, we divided the participants into three subsets of P: ones who did not change their ratings P_n, ones who changed P_c, and ones who skipped the question P_s. Our primary objective is to test the distributional properties of rating tuples from participants in P_n compared to those in P_c.

To ensure that all participants in the set P_c had an opportunity to see the median and then react, we filtered this group using the timestamps. The median appears in the interface with an animation whose completion time varied between devices, so we set a grace period of 3 seconds before we categorized the participant into set P_c.

For consistency, we use the same notation to describe participants in the reference survey. We denote the set of reference survey participants as set R, and each participant is associated with a 3-tuple ($g_i[j]$, $m[j]$, $g_f[j]$). However, since the reference survey does not reveal the median $g_i[j] = g_f[j]$ and $m[j]$ is the hypothetical median of the prior participants (which is not shown).

4. ANALYSIS PHASE

In the analysis phase, we determine whether social influence bias is statistically significant by analyzing spread of ratings around the median for the participants that changed their ratings. There are three principle challenges in testing this hypothesis. The first challenge is that parametric significance tests comparing two sample means such as the two sample t-test and z-test are known to perform poorly for multimodal and discrete distributions. Another significance test that is commonly applied to compare spreads of distributions is the F-test, which is also known to perform poorly for many non-normal distributions [20]. Furthermore, this test is usually used to test the spread of data around the mean, which only in very special conditions, such as normal distributions, aligns with the median which is the parameter of interest in the CRC. The discreteness of our data leads to multi-modal distributions which are not optimal for these testing methods.

The second challenge is that there is a natural tendency for ratings to concentrate around the median even without a bias. Consider the following participant behavioral model. Suppose that participants are not accustomed to a slider-based input. We can model the first rating that the participant leaves as uniformly randomly anywhere on the slider. As the participant begins to understand how to use the slider, their use becomes more accurate, ultimately settling on a rating from our observed distribution of final ratings. This model, the first rating is uniformly random and the second rating is a sample from the observed distribution, would result in a strong regression towards the median; even if there is no causal link with seeing the median.

Finally, the median m_i changes as ratings arrive and thus can be different for each participant. The median rating is calculated over all prior participants and thus is dependent on when the participant submitted their first rating. In practice, the median will eventually converge for a large number of participants, but it would be incorrect to measure concentration around a final median.

To address these three challenges, we propose a nonparametric model based on the Wilcoxon statistic to test the hypothesis that the group of participants that changed their ratings are more tightly centered around the median value that those participants observed. Our tests compare absolute deviations around the median for P_n, P_c, and R; which, as a relative comparison, controls for the natural tendency for ratings group around the median. Furthermore, it is more robust to the effects of alternate models such as the one described in our second challenge in comparison to a direct test of correlation (see Section 6.2.1).

4.1 Non-parametric Significance Test

Recall that P_n is the set of participants that did not change their ratings and P_c be the set of participants that changed their ratings. We define a set X_c, X_n of absolute deviations from the observed median of the final rating for each group:

$$X_c = \{|m[j] - g_f[j]|\} \ \forall j \in P_c \quad (1)$$

$$X_n = \{|m[j] - g_f[j]|\} \ \forall j \in P_n \quad (2)$$

For the purposes of hypothesis testing, we ignore the sign of the deviation. However, in Section 5, where we build a predictive model for the changes, we include the sign.

Now, for the set X_c, we calculate the Wilcoxon rank-sum statistic. We assign a rank to each of the absolute deviations in the union set $\mathbf{X} = X_c \cup X_n$ (ie. the largest change has rank 1 and the smallest has rank $|X_c \cup X_n|$. For X_c, we sum the ranks of the deviations within its set:

$$W_c = \sum_{j \in P_c} R_j \quad (3)$$

The *Null Hypothesis* is that absolute deviations in X_c are the same size as X_n. Under this null hypothesis $median(X_n) = median(X_c)$, the ranks will be evenly distributed between each group. Therefore, the null expected value and variance of W is:

$$\mathbb{E}(W) = \frac{(|\mathbf{X}|+1) \cdot |X_c|}{2} \quad (4)$$

$$var(W) = \frac{(|\mathbf{X}|+1) \cdot |X_c| \cdot |X_n|}{12} \quad (5)$$

For the significance level α, we can test the probability that our calculated W_c comes from the null distribution. In other words, the test calculates the probability that a random subset of users (ignoring the categorization P_n and P_c) can have the observed difference in rank-sum values. A significant result means that for the participants that changed their ratings the changed changes are more tightly centered around the median they observed. For many distributions, the Wilcoxon statistic is more robust as it uses ranks rather than the actual values, making it more resilient to outliers. Even in the case where the data is normally distributed, the optimal condition for the t-test, the relative efficiency of the Wilcoxon rank-sum statistic compared to the typically used t-statistic is $\frac{3}{\pi} = 95.4\%$. We tradeoff a small amount of efficiency in the normally distributed case, for increased efficiency and robustness in many non-normal distributions (eg. exponential $3\times$ more efficient). Recommender system data is almost always collected from discrete inputs which are usually not normally distributed.

The same analysis can be used to test X_c against the absolute deviations in the reference survey X_r

$$X_r = \{|m[j] - g_i[j]|\} \; \forall j \in R \quad (6)$$

or for initial vs. final ratings in the change group X_c':

$$X_c' = \{|m[j] - g_i[j]|\} \; \forall j \in P_c \quad (7)$$

4.2 Quantifying Concentration of Ratings

In addition to testing social influence bias, we can also estimate by how much the absolute deviations differ. The Wilcoxon statistic can be inverted to estimate a most likely *shift parameter* Δ, that is a shift Δ in the distribution of absolute deviations X_c that maximally aligns them with X_n. In other words, $X_c + \Delta$ is most supported by the null hypothesis (no social influence bias), or the distance from this hypothesis. An intuitive interpretation of Δ is that it measures how much our deviations have to be increased so that the no social influence bias hypothesis is the most likely conclusion. Since X_c is a set of absolute deviations, Δ tells us how much more concentrated X_c is than X_n around the observed medians. This parameter is relevant to the design of recommendation algorithms use similarity (eg. clustering or nearest neighbors), as it characterizes how much more on average are participants closer to the median.

We refer to [17] on the derivation of Δ and its confidence interval:

$$D = \{x_n[j] - x_c[i]\} \; \forall i, j \in X_n, X_c \quad (8)$$

$$\Delta = median(D) \quad (9)$$

5. BIAS MITIGATION

In our learning phase, we collect rating triplets ($g_i[j]$, $m[j]$, $g_f[j]$), and in our analysis phase, we determine whether the triplets exhibit statistically significant social influence bias. In the mitigation phase, we propose two models: correction model (infers the initial rating given a final rating and the median), and a prediction model (predicts final ratings given an initial rating and the median). Once trained, the correction model can be applied to correct final grades collected without the triplet (either historical or post-learning). The prediction model can be used to analyze properties of the social influence bias eg. are ratings above the median affected the same way as ratings below the median.

Previous work, suggests that social influence is not a homogeneous bias, namely, positive influences are different from negative influences. In Muchnik et al. [22], they found that when they positively treated posts with higher up-vote counts it lead to a significant increase in the likelihood of additional up votes (32% more likely). On the other hand, they argue negative treatments inspired correction behavior; where some participants wanted to correct what they felt was an incorrect score. They found that this also increased the likelihood of up-voting (88% more likely); as opposed to the conforming response which would be increased down-votes.

These results suggest that the effects of viewing median ratings can be non-linear and are very context/question dependent. Similar to the previous section where we applied non-parametric tests that did not make a strong assumption about the distribution of the data, we propose a information theoretic polynomial function search that does not make strong assumptions about the nature of the relationship.

5.1 Correction Model

Recall that $g_f[j]$ is the final rating for participant j, and $m[j] - g_i[j]$ is the difference between the median and the initial rating. We want to find a polynomial function f such that:

$$f(g_f[j]) \approx m[j] - g_i[j] \quad (10)$$

Let $f \in \mathcal{P}^k$ be a polynomial of degree k. The square loss of f, is the error in predicting $m[j] - g_i[j]$ from $f(g_f[j])$:

$$\mathcal{L}(X_c; f, k) = \sum_j ((m[j] - g_i[j]) - f(g_f[j]))^2 \quad (11)$$

For a given k, the best-fit polynomial minimizes this square-loss:

$$f_k^* = \arg\min_f \mathcal{L}(X_c; f, k) \quad (12)$$

For a given k, this problem can be solved with least squares. To search over the space of polynomial models, we apply a well-studied technique called the Bayesian Information Criterion (BIC) [8,26]. This technique converts the optimization problem into a penalized problem that jointly optimizes over the "complexity parameter" k. This penalty can be interpreted as bias towards lower degree models, in other words, an Occam's Razor prior belief. Cross-validation is an alternate method to empirically determine optimal model, and in practice, they give very similar results. BIC, however, is derived through maximum likelihood estimate and is not an empirical estimate so the learned model has a notion of optimality conditioned on the BIC prior belief.

Thus, we reformulate the optimization problem in the following way to incorporate the BIC penalty:

$$\arg\min_{f,k} |X_c| \log(\mathcal{L}(X_c; f, k)) + k \log(|X_c|) \quad (13)$$

The resulting optimal polynomial will tell how to correct a final rating to infer the initial one. Let q:

$$q(j) = m[j] - f(g_f[j]) \quad (14)$$

the predicted initial grade, and this value can be the input to our recommendation algorithm.

5.2 Applying the Corrections

There are two ways in which we can apply the correction model to existing recommender systems data. First, we can train our correction on all triplets, including ones that did not change, to get a correction that we can then apply to all ratings in the post-learning phase. The second way is to estimate the probability that a rating is changed, and if that probability is above a threshold α (eg. 50%) we can apply the correction. With the second way, the correction model is only trained on those triplets where the initial rating is different from the final one. To estimate this probability, we can apply a logistic regression model to predict whether or not a rating has been changed from all other ratings. Let $c(i,j)$ be 1 if participant j changed his or her rating for issue i and 0 if not. Our feature vector is the vector of all final ratings for that participant $v[j]_f = [g^1_f[j], ..., g^6_f[j]]$. Then, for the learned regression weights β, we can estimate the probability that $c(i,j) = 1$, using the logistic function:

$$P[c(i,j) = 1] = \frac{1}{e^{-\beta^T v[j]_f}} \quad (15)$$

We include results from both approaches in our experiments.

5.3 Prediction Model

For the prediction model, we make the dependent variable $m[j] - g_i[j]$ and the independent variable $g_f[j] - g_i[j]$. We apply the polynomial regression with the BIC optimization as before, and find an optimal function f such that

$$f(m[j] - g_i[j]) \approx g_f[j] - g_i[j] \quad (16)$$

f is a function of the difference between the initial rating and the median, that predicts the change in rating. This model allows us to reason about the nature of the social influence bias in the system. For example, if $|f(x)| > |f(-x)|$ for $x > 0$, we know that ratings above the median lead to a larger rating change. Additionally, $f'(x)$ tells us how the change varies as the observed difference with the median increases.

6. RESULTS

6.1 Dataset Description

The data for our case study was collected from the California Report Card between January 18th to April 20th. We also conducted an independent reference survey using SurveyMonkey's paid random panel system between March 8th and March 14th. As mentioned, ratings of six political issues were collected on a 13-point letter grade scale (A+,A,...,F) and for analysis we mapped these ratings linearly onto a scale from 0 to 1, with an F as 0 and A+ as 1. Participants also had the option to "skip" issues (not assign a grade). There were 1575 participants from the CRC and 611 participants from SurveyMonkey. Rating activity is summarized below.

Issue	No Change	Change	Skip	Median
CRC				
Obamacare	749	223	593	B (0.6667)
K12	849	172	544	C+ (0.5000)
College	923	139	503	C- (0.3333)
Immigration	693	105	767	C (0.4167)
Marijuana	881	118	566	C (0.4167)
Marriage Rights	929	105	531	B+ (0.7500)
Reference				
Obamacare	498	-	113	B (0.6667)
K12	561	-	50	C (0.4167)
College	573	-	38	C- (0.3333)
Immigration	375	-	236	C+ (0.5000)
Marijuana	498	-	113	C (0.4167)
Marriage Rights	554	-	57	B+ (0.7500)

For any given political issue, between 10% and 20% of those who assigned ratings registered a rating change. In all, 556 out of the 1575 CRC participants changed their rating at least once (Figure 3). We also found that the aggregate results of the reference survey matched the CRC nearly perfectly. On only two of the question (K12 and Immigration), we found a observed differences which were both less than a letter grade (+ or -).

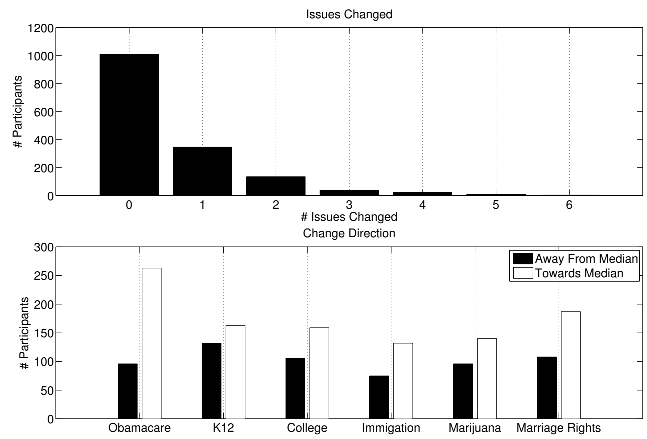

Figure 3: Among CRC participants, 65% changed none of their ratings, 22.0% changed one rating, 8.6% changed two, and 6.5% changed three or more. The lower figure inducates that majority of rating changes were towards the median.

6.2 Analysis

6.2.1 Correlation vs. Absolute Deviation

In Section 4, we argued that using correlation as a test statistic can lead to erroneous conclusions of social influence bias, and proposed testing the absolute deviations around the median. We ran an experiment to illustrate the problems of using correlation instead of absolute deviation. In this experiment, we iterated through the initial ratings each of participants in the change group P_c. For each rating, we randomly sampled a final rating from group P_n, the ones that did not change. In this model, since we sample final ratings from the no change group, we know that the social influence bias hypothesis is not true, since in distribution those who changed their ratings and those who didn't are exactly the same. However, when we calculate the correlation coefficient between $g_f[j] - m[j]$ and $g_i[j] - m[j]$, we find statistically significant correlations.

Issue	corr	p-val
Obamacare	0.709	5.2e-56
K12	0.659	4.73e-38
College	0.673	2.26e-36
Immigration	0.704	2.95e-32
Marijuana	0.689	1.42e-34
Marriage Rights	0.679	3.27e-41

There is a natural tendency for ratings to group around the median and the correlation coefficient does not account for this. However, if we measure the absolute deviation, we will find there is no statistically significant difference between the absolute deviations since they are the same in distribution.

6.2.2 Significance in CRC

Using the non-parametric test proposed in Section 4, we tested the hypothesis of whether rating changes led to significantly more concentration around the median. In our first experiment (Figure 4), we tested the absolute deviations of the CRC participants. We compared the group of participants that did not change their ratings to the group that changed their ratings. We found that while there were no statistically significant differences between the initial ratings of the two groups, the final ratings of the group that changed were statistically significantly more concentrated than both their own initial ratings and the ratings of the no change group. On average, the ratings were 19.3% closer to the median in the change group. The results of the hypothesis test for the set of participants who changed their ratings P_c and those who did not P_n are (we denote initial grades from P_c as i and final as f):

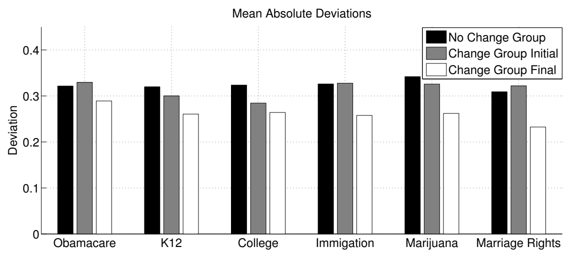

Figure 4: For those participants that changed their ratings, final ratings were significantly more concentrated around the median than their initial ratings. In addition, these ratings are more concentrated than the ratings for those who didn't change.

Issue	p-val(P_c vs. P_n)	p-val(i vs. f)
Obamacare	0.0286	0.0161
K12	2.1314e-06	0.0086
College	1.3033e-04	0.0415
Immigration	7.3456e-07	4.4170e-05
Marijuana	2.7549e-10	4.2560e-05
Marriage Rights	3.5946e-06	2.4644e-10

These results are consistent with social influence bias. When participants change their ratings, they are more likely to concentrate around the median. What is particularly surprising is that the two groups of participants P_n and P_c are very similar in terms of initial ratings, and the data suggests that a participant's susceptibility to social influence is not correlated with initial ratings.

6.2.3 Comparison to Reference Survey

In our second experiment (Figure 5), we apply the same testing procedure to compare the ratings from the CRC to to those in the reference survey. We compare absolute deviations of the group of participants who changed their ratings in the CRC against participants from the reference survey. The final ratings were 12.0% closer to the median in the CRC change group than in the reference survey. We also found that there was no statistically significant difference between the reference survey and initial ratings. The results of the hypothesis test for the set of participants who changed their ratings P_c and the reference group R are (we denote initial grades from P_c as i and final as f):

Issue	p-val(R vs. i)	p-val(R vs. f)
Obamacare	0.5386	0.0015
K12	0.8283	0.0097
College	0.1452	0.0091
Immigration	0.3765	1.1787e-04
Marijuana	0.7288	9.3111e-06
Marriage Rights	0.2478	0.0161

The results of our two experiments are consistent with social influence bias. We not only found that participants'

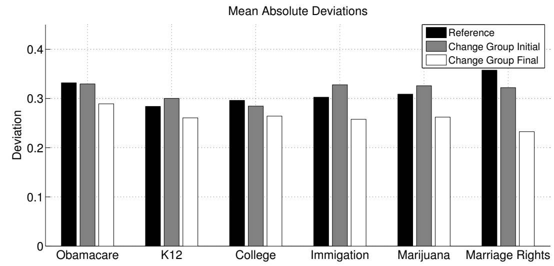

Figure 5: We found that final ratings were significantly more concentrated in the CRC compared to ratings in the reference survey. Similar to Figure 4, we found that there was no statistically significant difference between the reference survey and the initial ratings.

changed ratings were statistically significantly more likely to concentrate around the median, they were also more likely in comparison to the reference survey.

6.2.4 Estimating the Distance From the Null Hypothesis

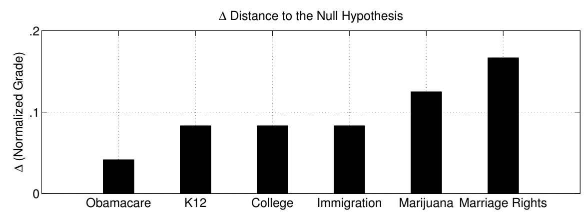

Figure 6: We calculate the shift-parameter Δ which is the distance from the null hypothesis. Over all issues, we found that ratings the average Δ was 0.0972 corresponding to a little more than a +/- grade.

We tested the hypotheses and conclude significant additional concentration of ratings around the median. In Section 4, we described how we could use the results of the hypothesis test to estimate the Δ parameter, which quantifies how different the hypothesis is from the null distribution (no social influence bias). In other words, how much would we have to spread our ratings around the median to negate the significant biasing result. We use Δ as a measure of social influence bias.

In Figure 6, we show the Δ estimates for each of the issues. For the issue about Marriage Rights, we find that parameter is largest at 0.1667. This means that all the final ratings for the Marriage Rights issue would have to be changed by 0.1667, corresponding to 2/3 of a letter grade eg. difference between B and A-, for us to conclude that there is no social influence bias in the dataset. For the other issues, the parameter was smaller indicating less of an effect of social influence bias. On average over all issues, the ratings were 0.0972 to a little more than a +/- grade.

6.3 Mitigation

6.3.1 Correction Model

We train the polynomial/BIC correction model proposed in Section 5, and evaluated it in terms of RMSE (Figure 8). We held out a random 20% of rating triplets and calculated the inference error in the correction model. We found that on average over all issues the RMSE was 0.1286 which

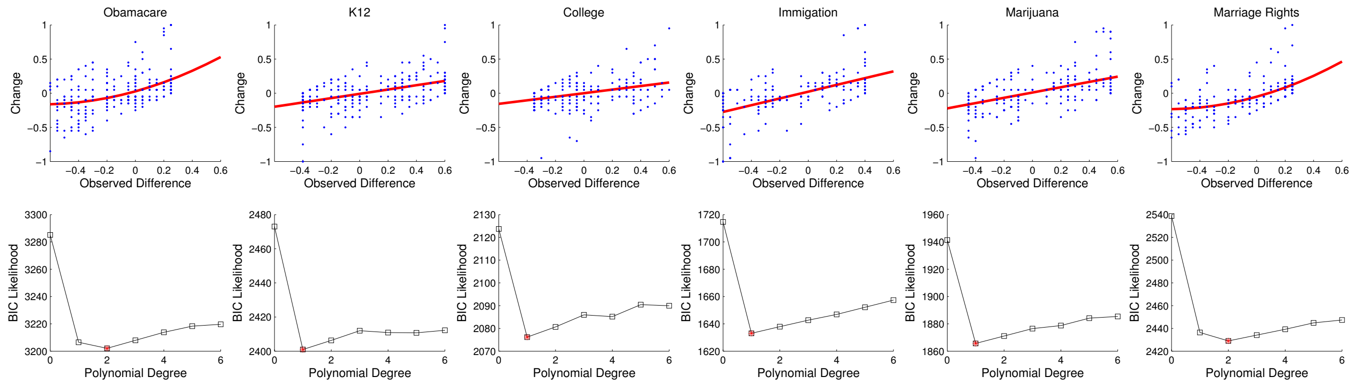

Figure 7: For the participants that changed their ratings, we plot the difference between their rating and the median (X-axis), and the change in their rating (Y-axis). We overlay the optimal polynomial model to represent the relationship $f(x) = y$. Below each plot, is the BIC objective function showing how we picked an optimal degree of polynomial.

corresponds to a little bit more than a + or - grade. We also measured the performance of the correction model by re-calculating the Δ, for the inferred initial ratings. A Δ of 0 means that the null hypothesis of no social influence bias is the most likely hypothesis, thus indicating perfect correction. We found that there was on average a 76.3% reduction in Δ.

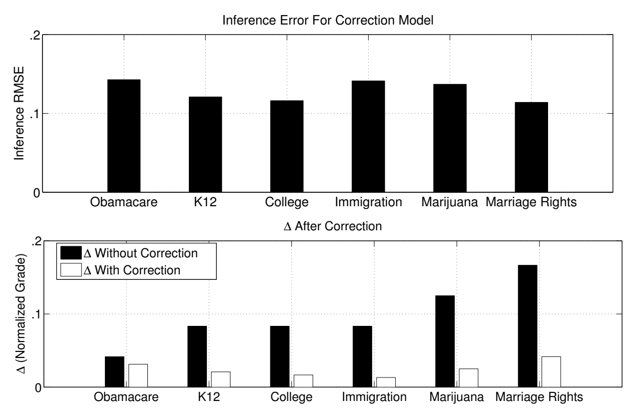

Figure 8: We measured the RMSE prediction error of the polynomial model. We found that we could predict changes in all of the issues with less than 2/3 of a letter grade RMSE error. In the lower figure, we applied this model to correct for the social influence bias and found that, on average, we could reduce the effects by 76.3%

6.3.2 *Classifying Final Grades As Changed*

In Section 5, we discussed how we could use logistic regression to estimate the probability that a rating has been changed. We applied logistic regression, as described in that section, and inferred which ratings were changed. In Figure 9, as is typically used to evaluate binary classifiers, we show the ROC plot of the logistic regression predictor. The prediction results were quite accurate with average AUC score over all issues of 0.8670. At the .50 probability threshold (classified as changed if the estimated probability is greater than 0.5), we achieved an average precision of 84.7% and a recall of 70.0%.

6.3.3 *Prediction Model*

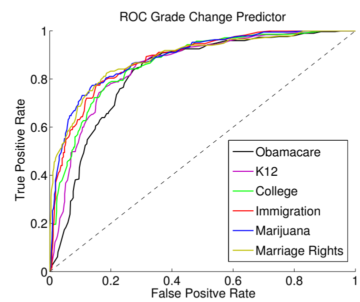

Figure 9: We show an ROC plot for the logistic regression estimate of the probability the rating has been changed. This plot shows the true positive rate (correct classifications) as a function of the false positive rate. We find that our prediction is quite accurate, substantially better than random (dashed line) with an average AUC score of 0.8670.

We applied the prediction from Section 5 and the results are shown in Figure 7. Our model search and optimization through the BIC discovered that for four out of the six issues, K12, College, Immigration, and Marijuana, the model was linear. This suggests homogeneity in positive and negative social influence effects for these issues. What this implies is that on average participants who rated above the median and below the median moved towards the median with the same magnitude. However, for Obamacare and Marriage Rights, we found that the relationship was quadratic. Interestingly enough, over the domain of changes, the learned quadratic function was "almost" linear, but with a steeper curve for ratings above the median. Participants who initially rated the state higher than the median had a more significant tendency to change downwards, in comparison to the upward tendency of those who rated less than the median.

6.4 How Many Training Examples?

It is important to note that our training set sizes were relatively small. Even with this small size, we were able to find an accurate correction model. The caveat is that only a fraction of participants will actually change their ratings during the learning phase. In Figure 10, we plot the optimality of the test error as a function of training set size for the correction model. We define the optimality percentage to be the ratio of the current test error to the best possible test error (test error using all 80% of the training set). We averaged the results over 1000 trials randomly picking a different 80% for training and 20% test. We found that a surprisingly few training examples could get a reasonably accurate model. For the linear models, we found that we could achieve greater than 95% optimality with only 25 examples. For the quadratic models, we required a little bit more data for the same optimality.

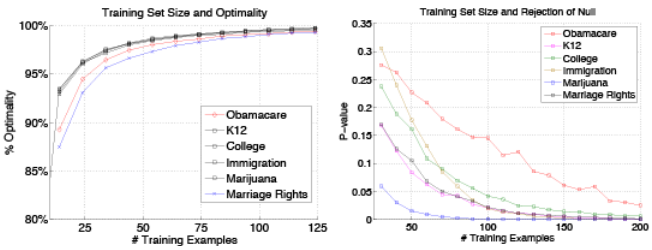

Figure 10: We find that we can train our correction model on less than 50 examples and get a model that on average performs only 5% worse than a model trained on the full training set. In the second plot, we plot p-values of our significance test as a function of training set size. We find for 5 out of 6 issues we could reject the null with less than 100 examples.

We can also look at the relationship between training set size and the analysis phase, where we are testing the statistical significance of the spread of the ratings around the median. Specifically, we can measure the average number of training examples needed before we can reject the null hypothesis at $p < 0.05$. In this dataset, we find that determining the significance of social influence bias requires more training examples than predicting its effects.

7. CONCLUSION AND FUTURE WORK

These results suggest that social influence bias can be significant in recommender systems and that this bias can be substantially reduced with machine learning. To apply this methodology to other recommender systems, a key question for future work is how is how to extend the approach to large item inventories and how much training data is required in such cases. One idea is to cluster/classify items into a small number of representative categories and train a model for each category. We believe that selecting an optimal set of items for training in this context may be posed as a submodular maximization problem. We are looking at applying this methodology to recommender systems in other domains (eg. movies) with alternative regression methods, such as Gaussian Process Regression and LOESS. We are also interested in performing more user studies where a false median is presented (as in the Asch experiments) and exploring methods to optimally classify participants as conformers and nonconformists. We would also like to study and quantify the role of social influence on textual data. textual data.

We would like to thank the UC Berkeley AMPLab and the UC CITRIS Data and Democracy Initiative for supporting this project. We also thank Brandie Nonnecke, Allen Huang, Camille Crittenden, John Scott, Tanja Aitamurto, Daniel Catterson, Matti Nelimarkka, Henry Brady, and Lt. Governor Gavin Newsom for their work on the CRC project.

8. REFERENCES

[1] J. Albors, J. C. Ramos, and J. L. Hervas. New learning network paradigms: Communities of objectives, crowdsourcing, wikis and open source. *International Journal of Information Management*, 28(3):194–202, 2008.

[2] S. E. Asch. Opinions and social pressure. *Readings about the social animal*, pages 17–26, 1955.

[3] S. E. Asch. *Studies of independence and conformity.* American Psychological Association, 1956.

[4] A. V. Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3):797–817, 1992.

[5] S. Bikhchandani and S. Sharma. Herd behavior in financial markets: A review. 2000.

[6] E. Bitton. A spatial model for collaborative filtering of comments in an online discussion forum. In *Proceedings of the third ACM conference on Recommender systems*, pages 393–396. ACM, 2009.

[7] R. Bond and P. B. Smith. Culture and conformity: A meta-analysis of studies using asch's (1952b, 1956) line judgment task. *Psychological bulletin*, 119(1):111, 1996.

[8] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach.* Springer, 2002.

[9] R. E. Burnkrant and A. Cousineau. Informational and normative social influence in buyer behavior. *Journal of Consumer research*, pages 206–215, 1975.

[10] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, pages 141–150. ACM, 2009.

[11] P. M. DeMarzo, D. Vayanos, and J. Zwiebel. Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics*, 118(3):909–968, 2003.

[12] U. M. Dholakia, S. Basuroy, and K. Soltysinski. Auction or agent (or both)? a study of moderators of the herding bias in digital auctions. *International Journal of Research in Marketing*, 19(2):115–130, 2002.

[13] S. Faridani. Using canonical correlation analysis for generalized sentiment analysis, product recommendation and search. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 355–358. ACM, 2011.

[14] S. Faridani, E. Bitton, K. Ryokai, and K. Goldberg. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1175–1184. ACM, 2010.

[15] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[16] J.-H. Huang and Y.-F. Chen. Herding in online product choice. *Psychology & Marketing*, 23(5):413–428, 2006.

[17] E. L. Lehmann and H. J. D'Abrera. *Nonparametrics: statistical methods based on ranks.* Springer New York, 2006.

[18] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, 2011.

[19] J. K. MacKie-Mason. Incentive-centered design for user-contributed content. 2012.

[20] C. A. Markowski and E. P. Markowski. Conditions for the effectiveness of a preliminary test of variance. *The American Statistician*, 44(4):322–326, 1990.

[21] S. Moscovici and C. Faucheux. Social influence, conformity bias, and the study of active minorities. *Advances in experimental social psychology*, 6:149–202, 1972.

[22] L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[23] T. Nathanson, E. Bitton, and K. Goldberg. Eigentaste 5.0: constant-time adaptability in a recommender system using item clustering. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 149–152. ACM, 2007.

[24] B. S. Noveck. Wiki-government. *Democracy: A Journal of Ideas (7)*, 2008.

[25] K. O'Hara. Transparency, open data and trust in government: Shaping the infosphere. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 223–232. ACM, 2012.

[26] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[27] R. Sipos, A. Ghosh, and T. Joachims. Was this review helpful to you? it depends! context and voting patterns in online content.

[28] W. Wood. Attitude change: Persuasion and social influence. *Annual review of psychology*, 51(1):539–570, 2000.

[29] H. Zhu, B. Huberman, and Y. Luon. To switch or not to switch: understanding social influence in online choices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2257–2266. ACM, 2012.