Temporal Mechanisms of Polarization in Online Reviews

Antonis Matakos
Department of Computer Science and Engineering
University of Ioannina
Ioannina, Greece
Email: amatakos@cs.uoi.gr

Panayiotis Tsaparas
Department of Computer Science and Engineering
University of Ioannina
Ioannina, Greece
Email: tsap@cs.uoi.gr

***Abstract*—In this paper we study the temporal evolution of review ratings. We observe that on average ratings tend to become more polarized over time. To explain this phenomenon we propose a simple model that captures the tendency of users for rating manipulation. Simulations with our model demonstrate that it is successful in capturing the aggregate behavior of the users.**

1. Introduction

The emergence of the Web 2.0 and social networking sites has enabled users to contribute content and actively participate in the online experience. One prime example of such active engagement is on review sites, where users can contribute opinions about products, places or people, and rate them on a numerical scale. Sites such as IMDB, Amazon, or Yelp play a critical role in capturing and shaping the public opinion. The public opinion is usually measured quantitatively by the average rating of the item, which is a concise way to aggregate the opinions of individual users into a single number.

However, although simple and effective, the average rating obviously hides information about the full spectrum of opinions for the item. This is why, many review sites often expose the full distribution of the ratings. Observing these distributions, a consistent pattern emerges in many diverse review sites. Most rating distributions follow a J-shape [4]: There is a large concentration of ratings in the top values (e.g., values 4 and 5), followed by few ratings in the middle values (e.g., ratings 3 and 2), and then a peak at the lowest rating values (e.g., rating 1). A typical example of such a J-shape distribution is shown in Figure 1a. At times the distribution takes the form of a U-shape (Figure 1b), where almost all the mass is divided between the two extreme ratings. Therefore, we can conclude that online ratings have an inherent degree of polarization. Opinions are not concentrated around the mean, but rather peak at the two ends of the spectrum.

Considerable effort has been devoted in analysing and modeling the temporal dynamics of review ratings. Driven by theories in psychology, a popular theory is that users exhibit a herding behavior when rating products (e.g., see [9]). However, this disagrees with the observed trends in the data.

(a) A J-shaped distribution (b) A U-shaped distribution

Figure 1: Two typical J-shaped and U-shaped distributions from Amazon.com

Wu and Huberman [10] study the evolution of book and movie review ratings, and they observe that the average rating declines, while polarization metrics such as variance increase. They hypothesize that this is due to self-selection bias: over time, only users that strongly disagree with the current mean rating will put the effort of writing a new review. Lorenz [7] analyzed user ratings for movies and found that they are double or triple-peaked. He proposed a model that approximates the distributions under the assumption that the histograms are discretized probability density functions of Levy skew alpha-stable distributions.

Another line of work has theorized that the polarization of the ratings can be explained by the discrepancy between the expected quality of the product created by reviews, and the realized quality after the product purchase [6], [8]. In a nutshell, the idea is that early positive reviews create high expectations that do not reflect the true item quality. Later reviews reflect the disillusionment of the users. Silva et al. [1] combine ideas from [10] and [6], [8] to explain polarization as the result of two parallel temporal processes in the data.

Our work extends and further confirms these findings. Key to our work is the assumption that users want to maximize their impact on the average when deciding on a particular rating score. While several models of user rating behavior have been proposed, to the best of our knowledge,

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA
978-1-5090-2846-7/16/$31.00

this is the first modeling effort which explicitly assumes that users manipulate the average when rating. More specifically, we make the following contributions:

- We study the temporal evolution of reviews. We demonstrate that (on average) reviews tend to become more polarized over time. This indicates that users do not exhibit a herding behavior towards a consensus, but rather they become more radical, adopting more extreme positions over time. Our results validate previous studies [10] showing the negative bias of existing opinions on the user behavior.
- We propose a simple model that could explain the behavior of the users, where we assume that users try to manipulate the public opinion towards their own. We demonstrate that this simple model can successfully capture some of the polarizing behavior of the users.

2. Temporal Evolution of Ratings

In this section we define a variety of metrics that characterize the aggregate behavior of ratings and we study their evolution over time over two datasets.

2.1. Datasets and Metrics

We consider two different datasets. For both datasets, we only consider items that have at least 100 ratings, giving us a sufficient sample of ratings to work with.

Amazon: This dataset, first used in [5], consists of a collection of review ratings for books on the Amazon site. The rating scores range from 1 to 5. The data is collected over the period of 10 years (5/96 - 5/06), and after pruning, it contains 1,334 items, and 291,567 ratings in total.

IMDB: This dataset, first used in [2], consists of a collection of review ratings for movies on the IMDB site. The rating scores range from 1 to 10. We only consider ratings that accompany a written review. The data is collected over the period of 6 years (01/08 - 09/13), and after pruning, it contains 464 items, and 123,394 ratings in total.

Both datasets that we consider are in the following form. We have a collection of items $\mathcal{I}$, e.g., a collection of movies. For each item x, we have a sequence of ratings $\{r_1^x, ..., r_{n_x}^x\}$, which are ordered by the time of their appearance. These ratings define a discrete *time-line* for the item x, where time-step k is defined by the appearance of the rating r_k^x. We will omit the superscript when immaterial.

We will now introduce the metrics for characterizing the temporal evolution of the ratings. For a given item, for each metric, we compute a value at time k based on the ratings $\{r_1, ..., r_k\}$. We consider the following metrics: (a) The *mean* at time k, defined as $\mu_k = \frac{1}{k}\sum_{i=1}^{k} r_i$; (b) The *variance* at time k defined as: $V_k = \frac{1}{k}\sum_{i=1}^{k}(r_i - \mu_k)^2$; (c) The *mean deviation* at time k, defined as:

$$\Delta_k = \frac{1}{k}\sum_{i=1}^{k} |r_i - \mu_{i-1}|$$

This measure captures how much a specific rating differs, on average, with the mean at the time of the rating; (d) The *kurtosis* at time k, defined as:

$$K_k = \frac{1}{k}\sum_{i=1}^{k} \frac{(r_i - \mu_k)^4}{\sigma^4} - 3$$

where σ is the standard deviation of the rating distribution. Kurtosis is the fourth central moment of the ratings distribution. It has been advocated by DiMaggio et al. [3] that the kurtosis is a good estimator of bimodality, which is a characteristic of polarization. Kurtosis takes high positive values for highly concentrated distributions, and low negative values for bimodal distributions. The Gaussian distribution has zero kurtosis.

Given a specific item, the values for any of the above metrics at each time step defines a time series for this item. We are interested in studying the average behavior of a collection of items. However, different items receive different number of reviews, so simple averaging is not possible. To circumvent this problem we compute the metrics at the first w-percentile of the ratings, for different values of w, for each item, and then aggregate these values to obtain an average time series.

For example, the mean rating of item x at the w-percentile is $\mu^x_{\lceil \frac{w*n}{100} \rceil}$, that is the mean of first $\frac{w*n}{100}$ ratings. Abusing the notation, we will denote this as μ^x_w. Given a collection $\mathcal{I}$ of N items we define the *average mean rating* at the w-percentile as

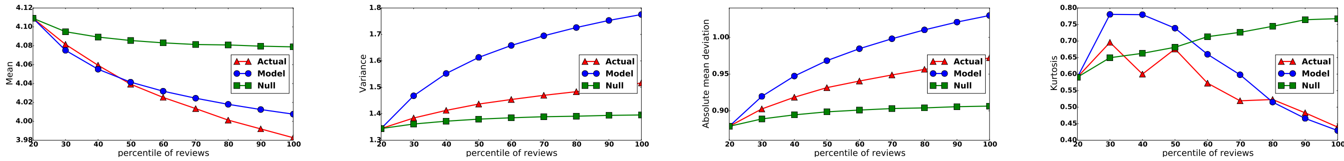
$$\overline{\mu_w} = \frac{1}{N}\sum_{x\in\mathcal{I}} \mu^x_w$$

Similarly, we define the *average variance* $\overline{V_w}$, *average mean divergence* $\overline{\Delta_w}$, and average kurtosis $\overline{K_w}$ at the w-percentile.

2.2. Temporal Evolution

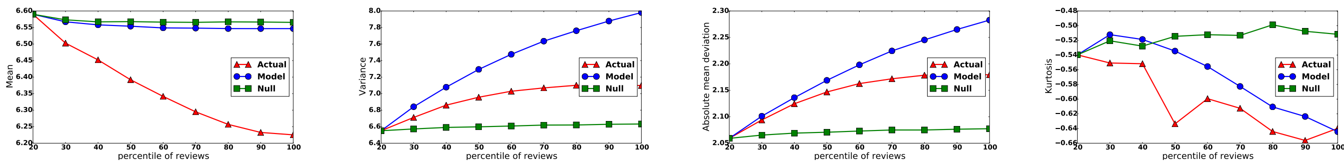
To study the temporal evolution of the ratings we compute the average values of the different metrics for increasing values of w. More specifically, we start with $w = 20$ and we increment w by 10, leading to time series of 8 points.

The red lines with the triangle markers in Figures 2 and 3 show our measurements for *Amazon* and *IMDB* respectively. We observe a consistent trend in both datasets: On average the mean rating drops, while variance and mean deviation increase and kurtosis drops, indicating an increase in polarization and variability. The following pattern seems to emerge in the behavior of the users: initially there is a positive outlook on the item, which over time is diluted by negative ratings. Our experiments clearly demonstrate the increase of polarization over time. We note the negative values of kurtosis for the *IMDB* dataset, indicating stronger polarization. Our observations agree with previous observations on the temporal evolution of ratings [10]. The experiments with kurtosis verify the polarization trend in the data.



(a) Average Mean (b) Average Variance (c) Average mean deviation (d) Average kurtosis

Figure 2: Average Metric Values over time for the *Amazon* dataset for actual and generated data. Model parameters: $p = 20\%$, $\alpha = 160$, and $\beta = k^{1.5}$.



(a) Average Mean (b) Average Variance (c) Average mean deviation (d) Average kurtosis

Figure 3: Average Metric Values over time for the *IMDB* dataset for actual and generated data. Model parameters: $p = 20\%$, $\alpha = 260$, and $\beta = 0.6 * k^{1.5}$.

3. A model for review polarization

From our experiments in Section 2 we observe that ratings tend to become more polarized over time, and that users are radicalized over time. We now propose a simple model that aims to explain this phenomenon. The intuition behind our model is that users react to the existing ratings, and they adopt more extreme opinions in order to "correct" the mean rating to be closer to their own. This behavior is actually documented in the review text of the reviews. For example in an IMDB review, a user writes: *"Is anyone else rating this a 1? I think I am going to so the 10 ratings can be balanced out."*.

To model this behavior we propose the following process for a user to produce their rating. Let r_k be the k-th rating in the sequence of ratings of a specific item, produced by some user u. We assume that the "true" opinion of the user u is not reflected by r_k. Instead, the true rating t_k of the user is drawn from a *base* distribution which reflects the true quality of the product. The "expressed" rating r_k is the product of a correction process performed by the user. Let μ_k denote the mean rating of the ratings up to r_k, including r_k. We assume that the user u incurs a *cost* for posting the rating r_k which is the following:

$$C(r_k) = \alpha(r_k - t_k)^2 + \beta(\mu_k - t_k)^2$$

This cost function has two parts. The first part captures the discordance of the expressed rating of the user with her own true opinion. The second part captures the correcting behavior of the user: she tries to correct the mean rating and bring it closer to her own opinion. This latter criterion leads to radicalization and polarization of the ratings. The user will try to adopt extreme positions in order to affect the mean rating as much as possible. The relative importance of the two factors is balanced by the parameters α and β.

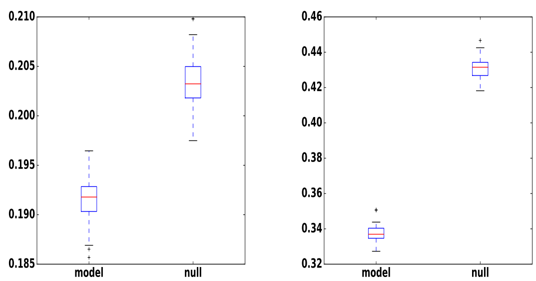
To minimize the cost, we take the first derivative and set it to zero. Solving for r_k, and using the fact that $\mu_k = \frac{(k-1)\mu_{k-1} + r_k}{k}$ we have

$$r_k = \frac{(\alpha k^2 + \beta k)t_k - \beta(k-1)\mu_{k-1}}{\alpha k^2 + \beta}$$

To test our model we need to define the *base distribution* from which we will draw the "true rating" of the user. Finding the true base distribution is hard, since it requires ground truth data where we solicit the unbiased opinions of the users. We make the simplifying assumption that the first ratings are unbiased, and thus capture to some extend the true quality of the item, so we use the first p-percentile of the ratings of the item to define the base distribution. To avoid zero probabilities, we smooth this distribution by adding one additional rating that we distribute equally among all rating denominations. Obtaining a better estimate of the true base distribution is an interesting problem for future study.

We now perform two different experiments. In the first experiment, we try to capture the trend in the evolution of the metrics using our model. The results are shown in Figures 2 and 3 for *Amazon* and *IMDB* respectively. The values achieved by our model are the blue lines with the circle markers in the plots. The values are averaged over 10 repetitions. For the *Amazon* dataset we set the parameters to $p = 20\%$, $\alpha = 160$, and $\beta = k^{1.5}$, while for the *IMDB* dataset we used $p = 20\%$, $\alpha = 260$, and $\beta = 0.6*k^{1.5}$. Note that β is a super-linear function of time k. This means that the effect of the correction factor in our cost function grows as time advances, intensifying the force of polarization. This agrees with the intuition put forth in [10] that as an item accumulates ratings, users are more motivated to post ratings that disagree with the average opinion. We observe that our model captures the increasing trend in both the variance and the mean deviation, and the decreasing trend in kurtosis. We also compare with a model that always draws from the base distribution (essentially, our model for $\beta = 0$). We refer to this as the *null* model, and it is shown in the green line with the square marker in the plots. As expected the behavior of the null model remains more or less constant over time. Interestingly the kurtosis value seems to increase.

In the second experiment, we compare the final rating distribution produced by our model for an item, to the actual distribution of the ratings for that item. To compute how close our estimated distribution is to the actual distribution, we compute the KL-divergence between the two distributions. We also compute the KL-divergence for the ratings produced by the *null* model. Figure 4 shows box plots for the distribution of the KL-divergence values for our model and the null model with parameters $p = 10\%$ and $\alpha = 1, \beta = 10$. We observe that our model achieves lower divergence for both datasets, and the two distributions are not overlapping at a confidence level of 95%.



(a) Amazon (b) IMDB

Figure 4: Box-plots for KL-divergence distribution for $p = 10\%$, $\alpha = 1$, and $\beta = 10$ for our model and the null model.

We experimented with a broad range of values and functions for the parameters p, α and β. Different parameter settings perform best for different metrics. We were not able to identify a parameter setting where we capture all metrics. Understanding and optimizing the parameters of the model in a principled way is an interesting problem for further research.

Gaussian baseline distribution. We also conducted a different experiment, where we assume that we know the actual true rating distribution of the ratings. In this stylized model the true rating distribution is a Gaussian centered around the observed mean rating, with standard deviation the observed standard deviation of the ratings. Our goal is to see if by applying our model on this baseline, we can create a distribution that is close to the observed distribution.

To test this, we use the Kolmogorov-Smirnov test that tests if two samples are drawn from the same distribution. For the parameters of our model, we used $\alpha = 1$ and $\beta = 30$ for the *IMDB* dataset, and $\beta = 100$, for the *Amazon* dataset, which we found experimentally to perform the best. At 0.05 significance level, the null hypothesis that the distribution produced by our model and the observed histogram are drawn from the same distribution could *not* be rejected for 57.7% of the movies, and 34.9% books respectively. Although, this does not confirm the null hypothesis, it is an indication for the validity of our model. Therefore, our results suggest that we can produce a distribution close to the true distribution for more than half of the items in the *IMDB* dataset. The results for *Amazon* are less impressive, but it is worth mentioning that the same test for the Gaussian without application of our model, rejects the null hypothesis for 95% of the books, demonstrating that clearly a Gaussian is not a good fit for our data.

4. Conclusion

In this paper we presented a study of temporal polarization in online reviews using data from IMDB and Amazon, and we proposed a simple model to explain this behavior. Simulations with our model show that it captures some of the behaviour exhibited by online reviewers. For future research, we want to devise a methodology for exploring the space of possible functions and values for the model parameters in order to fully understand the capabilities and limitations of our model. It would also be interesting to incorporate further temporal aspects in the model, in line with the findings of this and other similar studies.

Acknowledgments

This work has been supported by the Marie Curie Reintegration Grant project titled JMUGCS which has received research funding from the European Union.

References

[1] Sequential and temporal dynamics of online opinion. *Marketing Science*, 31(3):448–473, 2012.

[2] Q. Diao, M. Qiu, C.-Y. Wu, A. J. Smola, J. Jiang, and C. Wang. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *KDD*, 2014.

[3] P. DiMaggio and J. Evans. Have american's social attitudes become more polarized? *American Journal of Sociology*, 102(3):690-755, 1996.

[4] N. Hu, P. A. Pavlou, and J. Zhang. Can online reviews reveal a product's true quality?: Empirical findings and analytical modeling of online word-of-mouth communication. In *EC*. ACM, 2006.

[5] N. Jindal and B. Liu. Opinion spam and analysis. In *WSDM*, 2008.

[6] X. Li and L. M. Hitt. Self Selection and Information Role of Online Product Reviews. *Information Systems Research*, 19(4):456–474, 2008.

[7] J. Lorenz. Universality in movie rating distributions. *The European Physical Journal B - Condensed Matter and Complex Systems*, 71(2):251–258, 2009.

[8] A. Talwar, R. Jurca, and B. Faltings. Understanding user behavior in online feedback reporting. In *EC*. ACM, 2007.

[9] T. Wang, D. Wang, and F. Wang. Quantifying herding effects in crowd wisdom. In *KDD*, 2014.

[10] F. Wu and B. A. Huberman. Public discourse in the web does not exhibit group polarization. *CoRR*, abs/0805.3537, 2008.