ONLINE RATINGS: CONVERGENCE TOWARDS A POSITIVE PERSPECTIVE?

Yaonan Zhang[⋆] *Theodoros Lappas*[†] *Mark Crovella*[†] *Eric D. Kolaczyk* [⋆]

[⋆]Department of Mathematics and Statistics
Boston University
[†] Department of Computer Science
Boston University

ABSTRACT

Do online reviews reflect the true quality of products? Several articles, in both the popular press and the research community, have publicized that the average rating for top review sites is above 4 out of 5 stars. In this paper, we study the phenomena of review rating trends and convergence. We analyze data obtained from a popular restaurant review website, and present several models of increasing sophistication for the dynamics of the review ratings we observe.

Index Terms— online ratings, convergence, ordinal logistic regression, social influence

1. INTRODUCTION

Lots of review websites enable users to submit reviews to comment on the various aspects of a product. These reviews along with their ratings (e.g. usually scaled 1 to 5) become a major information source for subsequent users looking to make purchase decisions. In addition, brands that offer their products and services online have recognized the effect of reviews on their sales and reputation. Finally, third-party websites that host reviews on various products are continuously improving their review management platforms, in order to provide better service to both customer and brands. This effect of information diffusion through review websites is a digital version of word of mouth (WOM), as opposed to traditional off-line WOM which usually happens through acquaintances [1].

The average review rating of an item is the main piece of metadata provided by review websites. As important as they are, however, review ratings are not very well understood. Intuitively, the average rating is assumed to reflect the opinion of a representative reviewer. However, Hu, Pavlou and Zhang [2] examined the underlying distributions of the ratings of books, DVDs and videos from Amazon (http://amazon.com), and showed that the distributions are in fact bimodal. The bimodal distribution positions the average rating as a compromise of two groups of extreme opposite opinions, rather than an accurate representation of the true quality of an item. A relevant line of work tries to interpret the observed bimodal distribution via studying the intentions and mindset of review writers. The distribution is attached to a Brag-and-Moan Model, which assumes that users only write reviews when their experience with the product is either very good or very bad [2].

A number of earlier works have explored the effect of previously submitted reviews on the rating submitted by a new reviewer. Talwar, Jurca and Faltings [3] identify a user's rating as a partial reflection of the difference between the item's true quality and the user's prior expectation of quality, as inferred from previous reviews. The bias motivated by prior evaluations is shown to also be present in other review-based measures, such as the popular *helpfulness* measure [4]. Further, randomized experiments on social-news aggregation websites have shown that positive votes create positive herding effects, while negative votes are followed by a strong correction effect [5].

These findings motivate us to consider the connection between a potential positive bias and the well-documented observation that the average star rating on large review websites is very high, typically above four stars [6, 7]. In our work, we test the hypothesis that review ratings tend to converge towards an overall positive perspective, using data from the popular review website TripAdvisor (http://www.tripadvisor.com). Contrary to previous works that consider the distribution of the review ratings of all available reviews, we monitor this distribution over time, and test whether it tends to converge to a positive-dominated state. We also study the effect that the arrival rate of new reviews has on the average rating for an item. Our findings verify the upward trend of the average rating, as well as, in a nontrivial subset of the data, the connection between this trend and the arrival rate of new ratings.

The rest of this paper is organized as follows. In Section 2, we provide an initial, rudimentary analysis of the characteristics of review ratings in our data, aimed at establishing proof-of-concept. We find that while this analysis already is suffi-

This work was supported in part by AFOSR award 12RSL042 and NSF grant CNS-0905565.

cient to show that the data roughly confirms our hypothesis, it also leaves us further questions. In Section 3, we therefore present a more refined treatment of the problem, modeling the review ratings with an ordinal multinomial logistic regression. With this model we are again able to confirm the upward trend of ratings as time increases, now both in aggregate and, for a nontrivial subset of the data, at the level of individual restaurants. In Section 4, we next test to what extent the review arrival intensity – a natural and convenient proxy for the effects of WOM – contributes to the dynamics of the upward trend of ratings over time. Finally, some additional discussion and conclusions may be found in Section 5.

2. DATA AND A PRELIMINARY ANALYSIS

The context within which we explore the dynamics of online ratings is that of restaurant reviews. Specifically, we obtained restaurant reviews from the popular review website TripAdvisor (http://www.tripadvisor.com). Our data set has 1467 restaurants, consisting of all restaurants in NYC that were found to have at least 20 reviews. Reviews that did not contain opinions were ignored and did not count toward the 20. The time frame associated with our data is Oct. 2003 to Dec. 2012. For each restaurant, the information we use are the dates when the reviews were posted and the ratings, scaled as integers 1 to 5, that accompanied the reviews.

The question of whether there is a convergence towards a positive perspective in a collection of ratings can be usefully interpreted, from a statistical perspective, as asking whether the ratings have an increasing mean and decreasing standard deviation over time. In this section, we present an initial, rudimentary analysis of the restaurant review data, to see if the data confirms our hypothesis. Motivated by further questions raised by the results of this analysis, we then perform several more sophisticated analyses in subsequent sections.

Because it can be expected that the extent to which convergence in ratings is manifested will be influenced by the length of time over which reviews were posted, we divided the 1467 restaurants into six groups, corresponding to increasingly longer review periods. We defined the total length of a review period as $T_{\text{last}} - T_{\text{first}}$ (days), where T_{first} is the time of the first review in our dataset for a given restaurant, and T_{last} is the time of the last review. Group 1 corresponds to restaurants reviewed over the shortest period of time, and Group 6, the longest period of time. The distribution of the 1467 restaurants over these six groups is summarized in Table 1 below. Groups 3-5 have over 300 restaurants, whereas Groups 1, 2, and 6 have fewer (i.e., on the order of $50 - 150$).

Group Name	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	All
$T_{\text{last}} - T_{\text{first}}$ (days)	(0,600]	(600,1200]	(1200,1800]	(1800,2400]	(2400,3000]	(3000,3600]	(0,3600]
Restaurant Count	107	168	371	366	398	66	1476

Table 1. NYC review data (TripAdvisor)

Within each group of restaurants, we calculated the mean and standard deviation of reviews over a sequence of non-overlapping windows, each of 84 days in length. More specifically, within each group, all reviews that arrived in the first 84 days (starting from the time of each restaurant's first review) were pooled together, and the average and standard deviation of all rating in that window were calculated. Similarly, the same statistics were calculated for the next 84 days, and so on and so forth.

The results of this analysis are shown in Figure 1 and appear to strongly suggest, particularly for groups with longer review periods, that there is indeed an increasing mean and a corresponding decreasing standard deviation. We take these observed patterns as a rough confirmation of our hypothesis of the convergence of ratings toward a positive perspective in these data. However, a more refined treatment is clearly necessary, given the amount of oscillation still evident in these curves. Additionally, we wish to examine the robustness of our conclusions to the granularity of aggregation used here, i.e., over time and over restaurants. We pursue these issues in the following two sections.

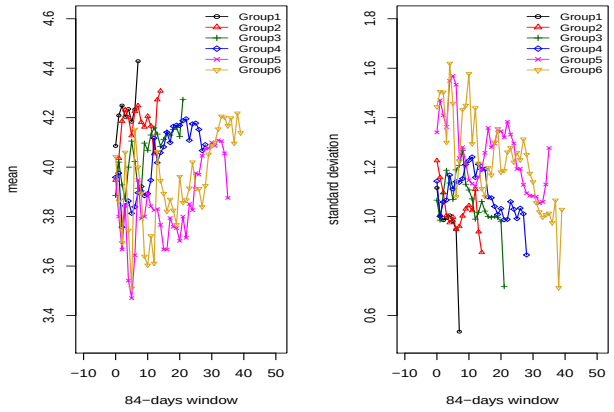

Fig. 1. Mean and standard deviation of restaurant review scores, in a moving window of time (84 days). Restaurants grouped according to total time being reviewed (see Table 1).

3. MODELING REVIEW RATINGS

Since the ratings have discrete integer values 1 to 5, it is natural to model them with multinomial logistic regression. Instead of modeling the probability mass function for each category (1 to 5), we model cumulative probabilities. Let Y_i be the rating at time t_i, where t_i indexes the i-th time window of some fixed length. Consider a model based on the cumulative response probabilities $\gamma_{ij} = \mathbf{P}(Y_i \leq j)$,

$$\log \frac{\gamma_{ij}}{1 - \gamma_{i1}} = \theta_j - \alpha t_i, \quad \text{for } j = 1, 2, 3, 4 \ , \qquad (1)$$

where α here is constant across response categories j. This model is called an ordinal logistic regression model [8, 9, 10]. It is also known as a proportional-odds model, because the ratio of the odds of the event $Y_i \leq j$ at t_1 and t_2 is independent of the choice of the category j. To fit these models we used the *polr* function in the R package *MASS*.

This model has a very nice latent variable interpretation [9], which is directly relevant to answering our question regarding rating convergence. Assume consumer opinion about a restaurant is actually a continuous variable Z, that $\epsilon = Z - \alpha t$ has the standard logistic distribution, and that there are thresholds $\theta_1, \theta_2, \theta_3, \theta_4$. If Z lies below θ_1 then a rating of 1 is given; if Z lies above θ_4 then a rating of 5 is given; and if Z is between θ_{j-1} and θ_j then a rating of j is given, for $j = 2, 3, 4$. This leads to a re-expression of our model as

$$P(Y_i \leq j) = P(Z_i \leq \theta_j) = P(Z_i - \alpha t_i \leq \theta_j - \alpha t_i) \quad (2)$$

$$= \frac{\exp(\theta_j - \alpha t_i)}{1 - \exp(\theta_j - \alpha t_i)} \ . \quad (3)$$

Thus when α is positive, the latent variable Z has an increasing mean as t increases.

With the understanding of this interpretation, we propose to evaluate and compare the following two models:

$$m_0 : \log \frac{\gamma_{ij}}{1 - \gamma_{i1}} = \theta_j, \text{ for } j = 1, 2, 3, 4 \ , \quad (4)$$

$$m_1 : \log \frac{\gamma_{ij}}{1 - \gamma_{i1}} = \theta_j - \alpha t_i, \text{ for } j = 1, 2, 3, 4 \ . \quad (5)$$

We fitted the linear model m_1 to each of the six groups data of data defined in the previous section. The fitted values of α were positive for all six groups, and statistically significant (comparing to the reduced model m_0) for all but groups 1 and 2 (these two consisting of restaurants having the shortest total review periods). The fitted mean and standard deviations are shown in Figure 2. They closely follow the trends we observed in Figure 1, which again confirms our hypothesis of convergence towards positivity in ratings. Moreover, the rate of change in the mean and standard deviation curves looks fairly consistent across groups.

We repeated the same analysis at the level of individual restaurants, to assess the extent to which the same increasing trend persisted at this level. Table 2(a) summarizes the results, again broken down by group. The first row is the total number of restaurants that have data in three or more categories of ratings. If a restaurant received only two (or fewer) rating values, it was not used for this particular analysis, as there was insufficient information to assess trend (i.e., manifesting in numerical instabilities during model fitting). Note that relatively few restaurants were excluded. The second row shows the proportion of restaurants found to have a statistically significant trend in their mean rating (i.e., model m_1 chosen over m_0). This number varies from about 10% to about 36% as the time exposure increases from Group 1 to Group 6. The third row shows, among those that have a linear trend, the proportion of restaurants having an *upward* trend. We see, therefore, that although only a third or fewer of the restaurants show a strong enough trend individually for us to detect, among those that did, it was overwhelmingly upward in nature.

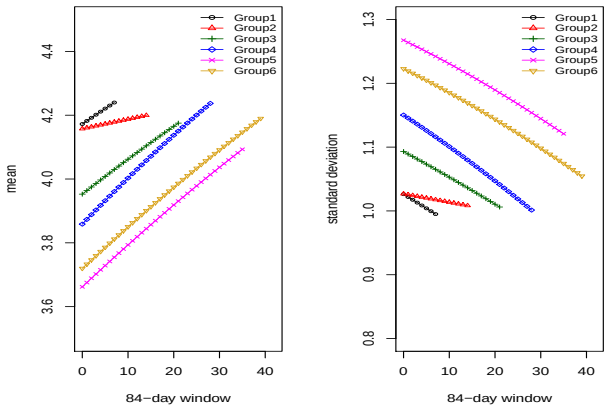

Fig. 2. Fitted Mean and standard deviation for the m_1. Data are grouped according to $T_{\text{last}} - T_{\text{first}}$.

Group #	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
N_{group}	102	165	363	365	397	66
$\frac{N_{m_1}}{N_{\text{group}}}$	10.78%	12.12%	12.12%	16.99%	22.67%	36.36%
$\frac{N_{m_1,\text{increase}}}{N_{m_1}}$	45.45%	35.00%	72.73%	95.16%	85.56%	91.67%

(a) With 84-day windows

Group #	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
N_{group}	105	165	363	365	397	66
$\frac{N_{m_1}}{N_{\text{group}}}$	10.48%	13.33%	12.40%	17.26%	23.17%	34.85%
$\frac{N_{m_1,\text{increase}}}{N_{m_1}}$	54.55%	36.36%	75.56%	95.24%	83.70%	91.30%

(b) With daily windows

Table 2. Comparison of models m_0 and m_1. Here N_{group} is the number of restaurants with ratings in three or more categories; N_{m_1} is the number of those restaurants for which m_1 is judged better than m_0, based on an analysis of deviance; and $N_{m_1,\text{increase}}$ is the number of the latter for which the mean is increasing (i.e., the estimated α is both significant and positive). In (a), three restaurants which have reviews in less than three windows are excluded from Group 1.

We examined the robustness of these conclusions to our choice of time resolution. As an extreme, we re-ran the analysis using windows of one day in length, rather than 84 days. The results are summarized in Table 2(b), where it can be seen that the numbers are very similar to those in Table 2(a).

4. ACCOUNTING FOR POPULARITY

In previous sections we have seen the upward trend of the review ratings as time increases. This motivates us to consider a simple mechanism that might be argued to push the average

rating higher over time, with an eye towards assessing the extent to which this mechanism explains some or all of the observed upward trend in our data.

Intuitively, we expect that a popular restaurant tends to get more positive reviews. This may be explained by the bandwagon effect, a group think behavior, and social influence on individual's perception of qualities, which has long been studied in economics and social science [11, 12, 13, 14], and recently in online social media [5]. Therefore, we propose to use the intensity of review postings as an indicator of how popular a restaurant is and to test, using an appropriately modified version of our models m_0 and m_1, to what extent the increase in review intensity explains increases in average rating.

We employ a two-stage procedure in our modeling, whereby we (i) use a non-parametric kernel-smoothing method for point process data [15] to estimate the review intensity, and then (ii) use ordinal multinomial logistic regression to model ratings. Specifically, letting $\lambda(t)$ be the review intensity at time t, and $\hat{\lambda}(t)$, the intensity estimated from the data, we compare the following two models:

$$n_0 : \log\frac{\gamma_{tj}}{1-\gamma_{t1}} = \theta_j - \beta\log(\hat{\lambda}(t)) \ , \tag{6}$$

$$n_1 : \log\frac{\gamma_{tj}}{1-\gamma_{t1}} = \theta_j - \beta\log(\hat{\lambda}(t)) - \alpha t \ , \tag{7}$$

for $j = 1, 2, 3, 4$. We note that the exact time of posting of reviews is not available to us, beyond the day of posting. Ties among the 'arrival time' of reviews can be broken through randomization, although this does not appear to affect our results.

Group #	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
N_{group}	105	165	363	365	397	66
$\frac{N_{m_1,\text{increase}}}{N_{\text{group}}}$	5.71%	4.85%	9.37%	16.44%	19.40%	31.82%
$\frac{N_{n_1,\text{increase}}}{N_{\text{group}}}$	4.76%	3.63%	8.82%	13.70%	7.56%	16.67%
$\frac{N_{n_1,\text{increase}}}{N_{m_1,\text{increase}}}$	83.33%	75.00%	94.12%	83.33%	38.96%	52.38%

Table 3. $\frac{N_{n_1,\text{increase}}}{N_{\text{group}}}$ is the proportion of restaurant where n_1 beats n_0 and still have a positive trend in time.

The results of our analysis, and the comparison of those results to our previous analysis, are summarized in Table 3. The second row of this table is the proportion of restaurants originally found to have an upward trend in time (calculated by multiplying the second row of Table 2(b) by the third row of Table 2(b)). The third row of Table 3 is the proportion of restaurants that have an upward trend in time even after accounting for the effect of the rating intensity λ. Finally, the fourth row of this table is the third row of Table 3 divided by the second row of Table 3. Rating intensity has higher explanatory power of the upward trend when the number in the fourth row is lower. We see that this number is much lower for Group 5 and Group 6 than for Groups 1-4, which means intensity has explained a lot of the upward trend in these two groups. These findings support our hypothesis that a popular restaurant is more likely to attract better ratings, hinting at latent group think/social influence factor. However, at the same time, our results indicate that there is still a substantial fraction of the upward trend in ratings that is not explained by our proxy for popularity.

5. DISCUSSION

The quality of the review ratings as well as the psychological/sociological reasons behind reviewer behaviors have become an interesting topic as the internet dramatically facilitates the effect of WOM among users. It is observed by practitioners and also mentioned in a few research articles that many large review websites have very high average ratings. We thus hypothesized that the distribution of the review ratings converges to a positive perspective as time increases. In this paper, we quantify this phenomenon first by plotting the rough characteristics of the ratings, then by subtler statistical modeling with ordinal logistic regressions. We found evidence that the ratings have an upward trend in time. This discovery is potentially a confounding effect between popularity and longevity, however, from the average ratings in Figure 1 we see no obvious selection bias toward grouping the worst(best)-quality restaurants into Group 1(6). Finally we tried to explain the trend using the popularity of the restaurants with the review intensity as a proxy for the popularity. We found interesting results that could support the group think/social influence hypothesis.

The latter was motivated by our assertion that the high influx of reviews that characterizes a popular restaurant is very likely to introduce numerous positive ratings, thus and lead to a converged positive state for the observed average. Looking forward, however, ideally a more nuanced approach to the joint modeling of ratings and review intensity would allow for dynamic feedback between the two, rather than intensity serving only as an explanatory variable to ratings, as in our analysis. For example, the connection between the arrival rate of new reviews and the increased average rating can also be explained if one considers the true quality of the reviewed items and the nature of ranking mechanisms on review websites: a high-quality item will eventually be discovered by users and receive the praise (and high ratings) that it deserves. A higher influx of reviews will give more visibility to the item, since it will be ranked higher by the website, and users also tend to gravitate toward frequently-reviewed items. As a result, the product's high quality will emerge faster, and become reflected on the high average rating. In other words, the increased arrival rate leads users (and thus new reviewers) to popular items of well-tested quality, that end up receiving even more positive reviews and increasing their average rating. Models capturing these types of dynamics are decidedly more challenging to define and to fit in a stable fashion, necessarily making their beyond the scope of the present work.

6. REFERENCES

[1] Frank M Bass, "A new product growth model for consumer durables," *Management Science*, vol. 15, pp. 215–227, January 1969.

[2] Nan Hu, Paul A Pavlou, and Jennifer Zhang, "Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication," in *Proceedings of the 7th ACM conference on Electronic commerce*. ACM, 2006, pp. 324–330.

[3] Arjun Talwar, Radu Jurca, and Boi Faltings, "Understanding user behavior in online feedback reporting," in *Proceedings of the 8th ACM conference on Electronic commerce*. ACM, 2007, pp. 134–142.

[4] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian 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*. ACM, 2009, pp. 141–150.

[5] Lev Muchnik, Sinan Aral, and Sean J Taylor, "Social influence bias: A randomized experiment," *Science*, vol. 341(6146), pp. 647–651, August 2013.

[6] Judith A Chevalier and Dina Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of marketing research*, vol. 43, pp. 354–54, August 2006.

[7] G. A. Fowler and J.D. Avila, "On the internet, everyone's a critic but they're not very critical," *Wall Street Journal*, 2009.

[8] Peter McCullagh, "Regression models for ordinal data," *Journal of the royal statistical society. Series B (Methodological)*, pp. 109–142, 1980.

[9] Peter MacCullagh and John Ashworth Nelder, *Generalized linear models*, vol. 37, CRC press, 1989.

[10] Ann A O'Connell, *Logistic regression models for ordinal response variables*, vol. 146, Sage, 2006.

[11] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of political Economy*, pp. 992–1026, 1992.

[12] Mariano Tommasi and Kathryn Ierulli, *The new economics of human behaviour*, Cambridge University Press, 1995.

[13] Irving Lester Janis, *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*, Cengage Learning; 2 edition, 1982.

[14] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing, "How social influence can undermine the wisdom of crowd effect," *Proceedings of the National Academy of Sciences*, vol. 108, no. 22, pp. 9020–9025, 2011.

[15] Peter Diggle, "A kernel method for smoothing point process data," *Applied Statistics*, pp. 138–147, 1985.