On the Volatility of Online Ratings: An Empirical Study

Christopher S. Leberknight , Soumya Sen, and Mung Chiang

Princeton University, Department of Electrical Engineering, Olden Street, Princeton, NJ 08544
{csl, soumyas, chiangm}@princeton.edu

Abstract. Many online rating systems represent product quality using metrics such as the mean and the distribution of ratings. However, the mean usually becomes stable as reviews accumulate, and consequently, it does not reflect the trend emerging from the latest user ratings. Additionally, understanding whether any variation in the trend is truly significant requires accounting for the volatility of the product's rating history. Developing better rating aggregation techniques should focus on quantifying the volatility in ratings to weight or discount older ratings. We present a theoretical model based on stock market metrics, known as the Average Rating Volatility (ARV), which captures the fluctuation present in these ratings. Next, ARV is mapped to the discounting factor for weighting (aging) past ratings and used as the coefficient in Brown's Simple Exponential Smoothing to produce an aggregate mean rating. This proposed method represents the "true" quality of a product more accurately because it accounts for both volatility and trend in the product's rating history. Empirical findings on rating volatility for several product categories using data from Amazon further motivate the need for the proposed methodology.

Keywords: Consumer confidence, e-commerce, decision support, online ratings, reputation systems.

1 Introduction

The effect of the current downturn of the U.S. Economy on consumer confidence can be observed by a close examination of the new shape in the demand and utility of e-commerce systems. Consumers are, now more than ever, looking for ways to minimize their expenses and e-commerce has moved from novelty and convenience to necessity. This has resulted in the growth of new technologies and revenue from applications such as mobile commerce (m-commerce) and location-based services. The combined revenue of e-commerce and m-commerce accounts for a large volume of the overall online retail market. Forrester Research estimates that both US and European online retail (representing 17 Western European nations) will grow at a 10 percent compound annual growth rate from 2010 to 2015, reaching $279 billion and €134 billion, respectively, in 2015.

Many e-commerce services like eBay and Amazon, as well as some of the newer online shopping applications, such as Yelp and Groupon, all share similar

characteristics and advantages over conventional retail stores in providing greater convenience, choice, and customer feedback. Almost all e-commerce systems offer potential buyers with easy access to product ratings from other buyers. The widely held belief is that virtual trust in these systems can be achieved when many consumers provide a similar rating for the product, independent of each other. This concept has been captured in phrases such as "wisdom of the crowd" which suggests that wisdom, and in this particular case quality, can be determined based on how many consumers share the same rating for the product. However, online ratings suffer from subjectivity bias and the arithmetic mean rating is often not the most accurate metric to base purchase decisions. Hu et al., [12] reported that online rating distribution for most products are bimodal (J-shaped) because of the "brag-and-moan" phenomenon among reviewers. Recent research also suggests other aggregators such as the weighted mean, median and mode may portray a more accurate picture of product quality [1].

While pursuing the identification of robust metric is undoubtedly useful, our focus in this paper is to understand and quantify the extent of volatility inherent in the user ratings of current systems. Since product ratings may be one of the strongest predictors of a consumer's decision to purchase a product online, providing such additional information on average volatility and latest rating trends can significantly influence consumer confidence. Using quantitative ideas from the stock market, we introduce a model to estimate rating volatility and use it to analyze an Amazon dataset with about 16,500 ratings across product categories of movies, books, fashion, and electronic goods. We examine the volatility in ratings that potential buyers of such goods are exposed to as they browse through the ratings from other reviewers, and then propose a way to potentially improve consumer's perception of a product's true quality.

2 Related Work

Research on online ratings has primarily focused on understanding various economic and social aspects, such as the impact of online ratings on sales [4],[6],[7],[11] and methods to increase trust and reputation [8],[10].

Besides the economic and social dimensions, researchers have also worked on the developing rating systems based on new quantitative metrics that may provide deeper insight and credibility for consumer ratings. Jøsang and Ismail [9] proposed a new rating metric called the beta reputation system that uses beta probability density functions to combine feedback and derive reputation scores. They also demonstrate an aging mechanism to account for the weighted difference between old and new reviews. Optimization techniques for new rating metrics have also been proposed [3]. In addition to the investigation of new metrics, other researchers have suggested simple transformations for displaying the quality of a given product. While rating feedback on product quality is most commonly aggregated using the arithmetic mean, recent research suggests other aggregators such as the weighted mean, median and mode may portray a more accurate picture of product quality [1]. Utilizing a different

aggregator such as the median was further substantiated through the examination of feedback bias [2].

Our paper complements these earlier works by empirically studying the volatility of ratings for several categories of products sold on Amazon. Such an approach can help in enhancing the reliability of existing rating systems as well as improving consumer confidence through better representation of the rating data. In our future work we intend to extend this study along the dimensions of aggregation granularity, stabilization, and rating prediction.

3 Model and Empirical Results

Many online rating systems today use a naïve averaging method for its product ratings (not to be confused with ranking), the simplest of which is an equally weighted average. But this is not necessarily a good metric for products that exhibit some clear trends. For example, consider the diagram shown on the left side of Fig. 1. It shows that there is a clear trend that the new ratings for the product are higher than the older ratings, possibly due to a substantial quality improvement. But averaging out the ratings for this product will fail to inform the buyers about this trend. Consequently, one could argue that discounting the older ratings (i.e. aging) to put greater weightage on the newer ratings would help to account for this latest trend. While that inference is correct, it is not clear whether aggressively discounting older ratings is always the right approach; for example, consider the product rating pattern shown on the right side of Fig. 1. The ratings show a high variability in the user's perception of the product's quality (e.g., because of some undetected flaw in the product). Therefore, in this case the older ratings should be given roughly equal weights in averaging instead of being discounted away. Hence, estimating a product's true rating would require a compromise between a simple averaging and a random walk model.

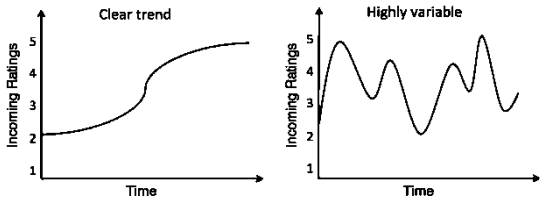

Fig. 1. (a) Left: ratings show a clear trend of improvement, (b) Right: rating is highly variable over time

In order to develop a prediction methodology that balances the need to account for variability in the ratings and the latest trends, we devise a three-stage strategy: First, we propose a metric called Average Rating Volatility (ARV) that captures the extent of fluctuation present in the ratings. Second, we define a map between ARV and the discounting (aging) factor to be used in weighting the past ratings. Third, we use this discounting factor as the coefficient for Brown's Simple Exponential Smoothing to

predict the mean rating that we believe is more 'accurate', given that it accounts for both past volatility and the latest trend. These steps are described next followed by an evaluation of the proposed rating strategy on the data gathered from Amazon for four different products categories (movies, books, cameras, and shoes).

3.1 Average Rating Volatility (ARV)

We introduce Average Rating Volatility (ARV) as a measure of the amount of variability inherent in the ratings that a product receives over time. It is calculated based on the mean rating value over non-overlapping intervals (windows) of N consecutive ratings. A time series of the product's rating fluctuation is thus obtained as the window moves. The counterpart to the idea of ARV in the stock market (Liu 2011) is known as the Volatility Index, denoted by VIX or AIV.

In this work, we compute the ARV based on a batch of 10 consecutive ratings, i.e. the window size N is set to 10. This size was chosen because most online portals and Amazon in particular, present viewers with a batch of 10 rated reviews on each page. Moreover, site visitors can view these ratings listed according to "newest first" option, in which case, the AVR metric reported here can be interpreted as a direct quantification of the volatility in ratings that a potential buyer is exposed to as he/she moves from one page to the other.

Let $R(t) = \{r_t, r_{t-1}, \ldots, r_0\}$ denote the original time series of a product's rating scores, and this series is divided into M non-overlapping time windows of size N, denoted by $W_1, W_2, \ldots W_M$. The Average Rating Volatility is then defined as the fractional change in the average rating values of two consecutive time windows of size N:

$$ARV(W_i) = \frac{|\langle R(W_{i+1})\rangle - \langle R(W_i)\rangle|}{\langle R(W_i)\rangle} \; ; \; i \in \{1, M\}\,. \qquad (\mathbf{1})$$

where $\langle R(W_i)\rangle$ is the average index value in window W_i, which is given by

$$\langle R(W_i)\rangle = \sum\nolimits_{k \in W_i} \theta_k r_k \quad ; \sum\nolimits_{k \in W_i} \theta_k = 1 \qquad (\mathbf{2})$$

θ_k's are the weights assigned to the ratings in window W_i, which in the simplest case is set to $\theta_k = 1/N$. A more sophisticated approach would entail setting the weights in accordance with the review of ratings (*e.g.* "helpful"/ "not helpful" comments left by other reviewers). The absolute value of the difference is used in (1) to capture the true magnitude of shift in public perception, which often shows a synchronized shift in one direction, upward or downward, depending on a sudden discovery of product flaw, introduction of rival products with comparable features etc. The ARV values thus give another time series, the area under which is a measure of the latent variability in the rating time series. But this area depends on the length of the time series data; hence we normalize it with respect to the total number of points in the ARV time series by computing the mean across all ARV windows. This measure, denoted as $\overline{ARV}$, represents the volatility of the product's ratings. For example, the

ARV of the 3rd generation iPod constructed from 1907 rating scores on Amazon is shown in Fig. 2. It is interesting to note that the volatility in the ratings have remained almost the same over time.

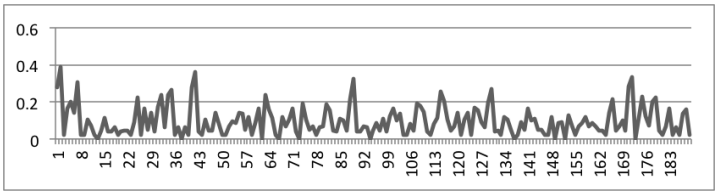

Fig. 2. Average rating volatility of iPod3 (32 GB) using 1,908 ratings from Amazon between 9/2009-12/2010

Table 1. $\overline{ARV}$ computed from 12,808 ratings across 10 commercially successful movies, with scores received on their Amazon DVDs between the release date and June 2005

Movies (Entertainment)	$\overline{ARV}$	**Number of ratings**
1. Independence Day (1996)	0.203208173	561
2. Titanic (1997)	0.172611210	1816
3. Saving Private Ryan (1998)	0.098727993	1196
4. The Big Lebowski (1998)	0.072623118	516
5. Armageddon (1999)	0.217262387	1171
6. American Beauty (1999)	0.124428681	1061
7. Fight Club (1999)	0.081965368	1276
8. The Matrix (1999)	0.094605664	2926
9. The Mummy (1999)	0.113956289	771
10. Gladiator (2000)	0.099715611	1514
Genre/Category, λ =	0.128 (12.8%)	Total = 12,808

Table 2. $\overline{ARV}$ for 5 critically acclaimed books computed between 1996 and 2005 using Amazon's rating data

Books (Arts & Culture)	$\overline{ARV}$	**Number of ratings**
1. One hundred years of solitude (G. G. Marquez)	0.095297574	424
2. Beloved (Toni Morrison)	0.146078200	557
3. Blindness (Jose Saramago)	0.098140184	251
4. Lolita (Vladimir Nabokov)	0.090344492	415
5. Satanic Verses (Salman Rushdie)	0.149658322	181
Genre/Category, λ =	0.116 (11.6%)	Total = 1,828

The ARV for the category of commercially successful movies around 1999 is shown in Table 1. The category $\overline{ARV}$, denoted by λ is computed to be 0.128, *i.e.* products of this category show on an average about 12.8% deviation from the mean rating of one page to the next on Amazon. Similar results on the average volatility in books, digital cameras, and women's shoes are reported in Tables 2, 3, and 4 respectively.

Table 3. $\overline{ARV}$ for a group of digital cameras computed between 2008 and 2011 using Amazon's rating data

Digital Cameras (Electronics)	$\overline{ARV}$	**Number of ratings**
1. Canon PowerShot SX20IS 12.1MP	0.087368	391
2. Nikon S8000-vibration reduction	0.178723893	552
3. Polaroid CZA-10011P	0.106465356	175
Genre/Category, λ =	0.124 (12.4%)	Total = 1,118

Table 4. $\overline{ARV}$ for popular women's shoe brands based on Amazon's rating data between 2006 and 2011

Women's Shoes (Fashion)	$\overline{ARV}$	**Number of ratings**
1. Easy Spirit Traveltime	0.056502532	150
2. UGG Classic Footwear	0.047895357	148
3. BearPaw Shearling Boots	0.086294498	201
4. Skechers Shape-Ups	0.12181668	186
5. Tamarac Slippers	0.121079089	120
Genre/Category, λ =	0.087 (8.7%)	Total = 805

3.2 Discounting Factor Function (DFF)

We now introduce a function that maps $\overline{ARV}$ (which provides a measure of the latent volatility in the product's ratings) to the coefficient with which older ratings need to be discounted. This coefficient is also referred to as the discounting (aging) factor, $0 \leq \alpha \leq 1$. A small value of α puts greater weights on past ratings while $\alpha \approx 1$ puts more the weight on the current rating in making future predictions.

When $\overline{ARV}$ is low (*i.e.,* similar average ratings across time windows), then the discounting factor should be kept large so that more weight is put on the current values, thus any sudden deviation in the trend gets reflected quickly. On the other hand, if $\overline{ARV}$ is high, (*i.e.,* the product's review are volatile), then the discounting factor is kept relatively low to average out any sudden fluctuations, thus responding only when there is a real trend. As shown in Figure 3, we choose a shifted sigmoid

function for this mapping between the discounting factor and the average volatility in a product's rating because it satisfies the above properties and makes the discounting factor steadily sensitive to changes around the product category's $\overline{\text{ARV}}$.

$$\alpha = \frac{1}{1 + e^{\eta(\overline{ARV} - \lambda)}} \tag{3}$$

where λ is the product's category $\overline{ARV}$ and can be computed as shown in Tables 1-4 of Section 3.1. The parameter η is a scaling constant to make α close to 1 when $\overline{ARV} \simeq 0$ and can be used to also control the sensitivity of the discounting factor to the rating volatility.

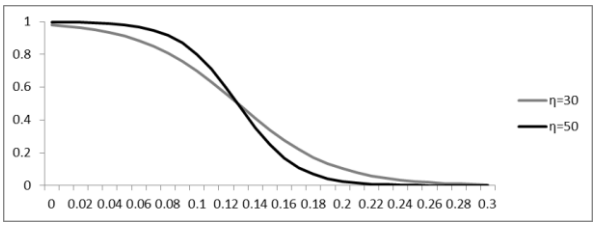

Fig. 3. Plot of aging factor versus (ARV) for λ=0.128

3.3 Simple Exponential Smoothing (SES)

SES is a method that discounts past data in a gradual fashion. For example, it ensures that the most recent rating gets a little more weight than 2nd most recent, and the 2nd most recent gets more weight than the 3rd most recent, and so on. This discounting is done by a "smoothing coefficient", which is the aging factor, α introduced in Section 3.2. The formula used by SES recursively to update the smoothed series as new ratings come in is given by:

$$R_t = \propto r_t + (1 - \propto) R_{t-1}; \qquad 0 \leq \propto \leq 1 \tag{4}$$

$$r_{t+1} = R_t;\ R_1 = r_0$$

Thus, the current smoothed value is an interpolation between the previous smoothed value and the current observation, where $\propto$ controls the closeness of the interpolated value to the most recent observation. Notice that if $\propto = 1$ the SES model is equivalent to a random walk model (without growth), while $\propto = 0$ makes it equivalent to a mean model, assuming that the first smoothed value is set equal to the mean. The average age of the rating data in the SES forecast is $1/\propto$, relative to the period for which the rating prediction is made. For example, when $\propto = 0.2$ the lag is 5 periods etc.

To summarize, our proposed methodology works as follows: for a given rating time series we first compute the mean ARV ($\overline{ARV}$), which is mapped to a discounting

factor that is used as the coefficient for the SES method to predict a better aggregate rating. Next this methodology is applied to two online products as shown in Fig. 4 and 5. In Figure 4, the evolution of Amazon's mean rating is depicted (solid black line) for the DVD of Armageddon, which does not reflect the dynamic changes in the real ratings from the users. By accounting for the high volatility in user ratings, a new ratings evolution (dashed black line) is constructed using the proposed approach. It can be observed that this ratings evolution is more responsive to changing user preferences while at the same time avoiding random short-term fluctuation trends that arise from user's subjectivity.

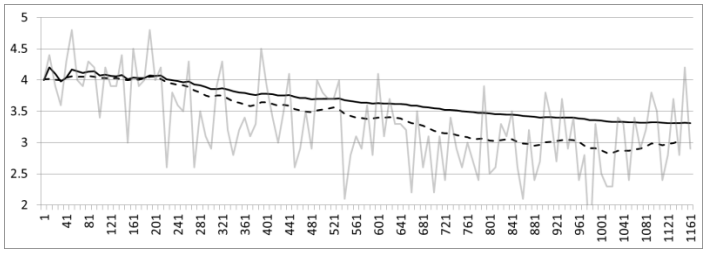

Fig. 4. Armageddon (11/1998-6/2005, 1171 ratings). Amazon's rating score is slow to change as it neglects rating volatility (Solid black: Evolution of the mean rating on Amazon, Grey: Evolution of the actual rating from batches of 10 consecutive ratings, Dashed black: Estimate of true rating that account for volatility using the proposed ARV approach)

In Fig. 5, the ratings for G. G. Marquez's famous novel lie in a much narrower range of values and hence have a low volatility in contrast to the movie Armageddon.

Fig. 5. G. G. Marquez's book One Hundred Years of Solitude (424 ratings). Amazon's rating score does not reveal the user's true score (Solid black: Evolution of the mean rating on Amazon, stabilizes at 4.5, Grey: Evolution of the actual rating from batches of 10 consecutive ratings, Dashed black: Estimate of correct rating while accounting for volatility using our ARV approach; the score goes down to 4 but Amazon users never get to see it)

Therefore, the prediction for the aggregate rating mostly follows the latest trend and is therefore more responsive to changes in the readership's taste or critical evaluation of the book in the larger social context. Similar results were observed for other categories of products that further strengthen the case for using this AVR-based rating strategy for online rating systems.

4 Conclusion and Future Work

This paper presents a theoretical model for a new rating mechanism which may provide greater insight into the reliability of existing ratings used to convey product quality. The three-tiered strategy of this model attempts to address issues of stabilization, prediction, and discounting. First, we introduce a metric called Average Rating Volatility (ARV) that captures the extent of fluctuation present in the ratings. Second, we define a map between ARV and the discounting (aging) factor to be used in weighting the past ratings. Third, the discounting factor is used as the coefficient for Brown's Simple Exponential Smoothing to predict the mean rating and future values. Lastly, the proposed model is evaluated using data gathered for different types of products from Amazon. A rating mechanism based on this model will provide a more accurate representation of "true" quality of a given product since it will account for both rating volatility and the emergence of long-term trends.

Acknowledgments. This work was supported in part by NSF CNS-0905086, CNS-1117126, and ARO W911NF-11-1-0036.

References

1. Garcin, F., Faltings, B., & Jurca, R. (2009). Aggregating reputation feedback. Proceedings of the International Conference on Reputation: Theory and Technology. 1(1): 62-74.
2. Jurca, R., Garcin, F., Talwar, A., & Faltings, B. (2010). Reporting incentives and biases in online review forums. ACM Transactions on the Web (TWEB). 4 (2):1–27.
3. De Kerchove, C. and Van Dooren, P. (2008). Reputation Systems and Optimization. SIAM News. 41(2).
4. Zhang, X. M., and Dellarocas, C. (2006). The lord of the ratings: How a movie's fate is influenced by reviews. Proceedings of the 27th International Conference on Information Systems (ICIS).
5. Liu, J., Tse, C. K., He, K. (2011). Fierce stock market fluctuation disrupts scale-free distribution. Quantitative Finance. 11(6): 817-823.
6. Bolton, G. E., Katok, E., & Ockenfels, A. (2004). How effective are on-line reputation mechanisms? An experimental investigation. Management Science. 50 (11): 1587–1602.
7. Duan, W., Bin G., & Whinston, A. B. (2008). Do online reviews matter? An empirical investigation of panel data. Decision Support Systems.
8. Resnick, P., Zeckhauser, R., Friedman, E., & Kuwabara, K. (2000): Reputation systems, Communications of the ACM. 43: 45–48.
9. Jøsang, A., and Ismail, R. (2002). The beta reputation system. In Proceedings of the 15th Bled Conference on Electronic Commerce.
10. Jøsang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. Decision Support Systems. 43(2): 618–644.
11. Hu, N., Pavlou, P. A., & Zhang, J. (2006). Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of Online word-of-mouth communication. Proceedings of the 7th ACM conference on Electronic commerce (EC).
12. Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. Communications of the ACM. 52(10): 144-147.