Effects of time-based Biases in Review Communities

Kevin Ho, Sean O’Donnell
CS224W Final Report

[0.0] Abstract

Prior work has used social biases to inform predictive models of evaluation-driven networks, but there still exist many opportunities to analyze and expand on frameworks for modeling the growth of review-based communities. In particular, we would like to more deeply explore the influence of the time of a review’s creation and the effect it has on other review metrics. In an observational study, we analyzed an Amazon Product Review data set to see if earlier reviews influenced successive, a behavior consistent with that of the Anchoring effect from work in social psychology. Our core analysis found that when there are a low number of reviews (~ 15), earlier reviews have a tendency of being closer to the mean than later reviews - while this implies that there may indeed exist an anchoring effect, further analysis found that for the same fixed bounds in review number, earlier reviews also were generally deemed more helpful. Through a comparative study with the similarly 5-star scaled Beer Advocate beverage evaluation network review data set, we conclude that certain divergent trends found in the Amazon Review dataset can be attributed to its review helpfulness mechanism, rather than an inherent anchoring effect based simply on time.

[1.0] Introduction

With the start of organized economies, came the necessity for being able to assign value to a product. Of course, the true “value” of a product can never be objectively quantified through currency, but any discussion of price and purchase needs a point from which to compare. In effect, whether it’s bargaining at a flea market or shopping online, the price always serves as an anchor from which to consider purchasing a product. Whereas the real-world and online markets frequently feature many of the same types of anchors (initial price, brand, venue), the online market differs significantly through the inclusion of robust, evaluation-driven models. Reviews add a completely new dimension of anchoring to the online market - ratings, easily accessible customer feedback, and review usefulness metrics have become a normal part of the online shopping experience, and each comes with its own potential as an anchor in consumer decision-making. More generally, the reviews included with any product have become yet another feature to consider in the purchase of products.

[1.1] Prior Work

A key piece of prior work was by C. Danescu-Niculescu-Mizil et al (2009) on the Amazon review network, where work was done to evaluate the relationship between helpfulness ratings and the average rating of a product. Based on social biases, the paper suggests 3 different cases which can model how helpful a given review will be rated, given the review’s relationship to the product’s average rating. Analysis done in this work concludes that the data is consistent with the conformity bias, due to the observation that reviews which differentiate the least from the computed average rating of a product are generally deemed the most helpful. Our work seeks to expand on this model in the dimensions of time and influence. In the case that there is a correlation between the time of rating and other product metadata, it is possible that earlier reviews have a first mover advantage, and therefore an disproportional influence on the average rating of the product. The Anchoring Effect as a common cognitive bias has been established through multiple social experiments (Ariely, Loewenstein, Prelec (2006); Saal, Downey, Lahey (1980)).

[2.0] Data

[2.1] *Amazon product co-purchasing network metadata*

Our studies investigate Amazon.com product metadata scraped in 2004, consisting of approximately 500,000 products, product information, as well as associated review data. The main data of interest from this set was in the evaluation component of the product network, specifically the review rating, number of votes for a rating, helpfulness, and the date of review. Within this set of products, 402724 products had at least one review. Along with these data points, we also intend on using the product category in order to see if there are any differences depending on the product type. In particular, we focus on the four following product categories with the largest number of reviewed products: Books (278,217), Music (83,659), Video(22,357), DVD(18,474).

[2.2] *BeerAdvocate Reviews*

During our analysis, we also used the BeerAdvocate.com review network, which consists of all reviews on the site from 1998 - 2011. This set of reviews was used for comparison against the Amazon data set as a baseline as both used a similar 1-5 scale to rate products, although BeerAdvocate allows for a finer resolution of review ratings through ratings such as 1.5, 2.5, etc. In contrast, Amazon.com only allows integer values. Furthermore, a key difference between the two datasets are the additional mechanics in assessing the validity of a review - Amazon.com extensively uses helpfulness ratings (the most helpfully rated review is the most visible review for any product), while such a mechanism is not present within the BeerAdvocate review site. Additional dataset statistics can be found in figure 1.0.

	Amazon	BeerAdvocate
Number of Reviews	1586614	7593244
Number of Products with Reviews	66055	402724
Number of Reviews Per Product	24	19
Average Variance within reviews for a Product	.1859	.7075
Average Rating	3.815	4.178

Figure 1.0: Aggregate dataset statistics for the Amazon product co-purchasing network and BeerAdvocate reviews.

[3.0] Hypotheses

Based on a model informed by the Anchoring effect, our initial hypotheses speculated that earlier reviews would deviate less from the computed average for a given product: the value perception of "good" products (above average), as they accumulate favorable reviews, will continue to increase, and vice versa for "bad" products, regardless of the actual inherent value of the product (which arguably, no objective metric exists for our Amazon data set of entertainment-based products). Therefore, we anticipated that this effect could be more strongly observed for products with a higher rating variance, as this high variance should indicate lack of an inherent "consensus" rating around which reviews would all tend to. Although our work has been performed through an observational rather than experimental study, we formally stated the following hypotheses prior to analysis:

H1: *Earlier reviews deviate less from the computed product's average rating than later reviews.*

H1.1: *Given H1, the deviation from the mean for earlier reviews is less for products with higher variance than it is for products with lower rating variance.*

As a consequence of the conformity effect observed in the Amazon dataset in prior work, we anticipated that the results of this study should also be influenced by this effect:

H2: *Earlier reviews are rated more helpful than later reviews.*

In the case that H2 is proven true through our analysis, it is unclear whether earlier reviews deviate less from the computed average because earlier reviews are rated more helpful, or whether the converse causality is true. To resolve this, we developed a final, third hypothesis:

H3: *The effects of the helpfulness mechanism in the Amazon network cause earlier reviews to be closer to the computed mean.*

Our analysis of H3 is primarily derived from a comparative study between the Amazon and BeerAdvocate dataset, as BeerAdvocate does not have a helpfulness mechanism in place.

[4.0] Methods and Results

[4.1] ***Effects of Time on the Deviation from Computed Product Average of Reviews***

As in the work done by C. Danescu-Niculescu-Mizil et al (2009), let the *computed product average rating* be defined as, “the average star rating computed over all star rating reviews of that product in our dataset.” Furthermore, let *date-rank* be defined as the rank of a given product review in the chronological ordering of reviews for a product. For example, the first review of a product has date-rank 1, and the third has date-rank 3. We check the first hypothesis that earlier reviews deviate less from the computed product average than later reviews by first ranking reviews for each product in order of increasing date, and then computing absolute difference, called the *deviation from computed average*, between computed product average and the star rating of each rating. The average deviation for a given date-rank is then computed by taking the average of the deviations for a given date-rank for ratings of the same date-rank from other products, where the total number of reviews, for that product are equal.

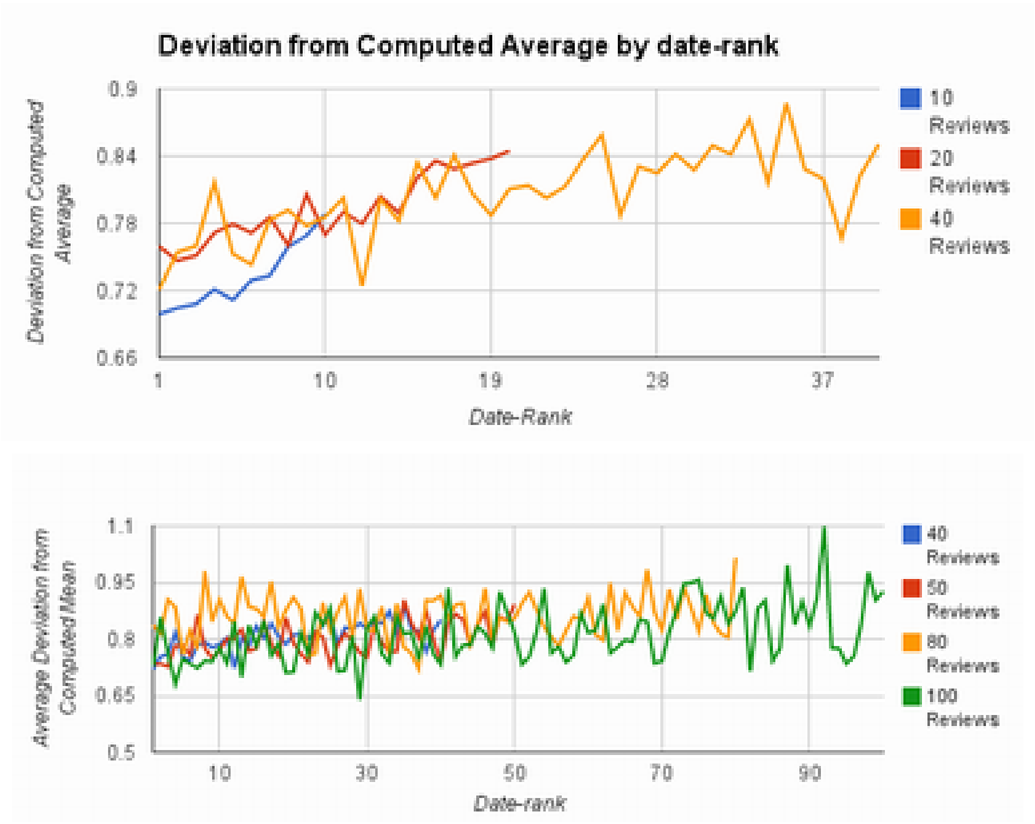


Figure 3.0: Average deviation from computed mean by date-rank. Each line represents the average deviation for a given rank for all products with the same number of total reviews.

Figure 3.0 shows the key values from the results of this analysis, which is over all products which have at least one review. We also performed analysis by the four largest product groups, as well as by different rating variances, but this produced data that presented similar results. This data is consistent with our hypothesis H1 and shows that lower date rank is correlated with a lower deviation from computed mean. However, this correlation is significantly stronger for products with less than 40 reviews; past this value, the correlation slowly diminishes to the point that the regression line is approximately zero, which can be seen from the simple linear regression results in figure 3.1.

Total reviews	10	20	40	50	80	100
N	8668	2962	955	595	284	225
β	.0095	.0048	.0022	.0016	.0003	.0012
α	.6801	.7428	.7613	.7587	.8500	.7501

Figure 3.1: Results of simple linear regression analysis on Amazon data for average deviation from computed mean, where $y = \alpha + \beta x$. N is the total number of products included in the average deviation calculation.

[4.2] ***Effects of Time on Helpfulness Rating***

While the data previously described validates H1, several confounding factors exist which could cause the above result, aside from an anchoring effect. In order to investigate such confounds, additional analysis was conducted to check whether the helpfulness rating of a review had a relation to the date-rank of a review. Our methodology was similar to that of the analysis for H1, with the exception that our measure was in this case the average helpfulness ratio, where helpfulness ratio is the number of helpful votes divided by the total votes for that review.

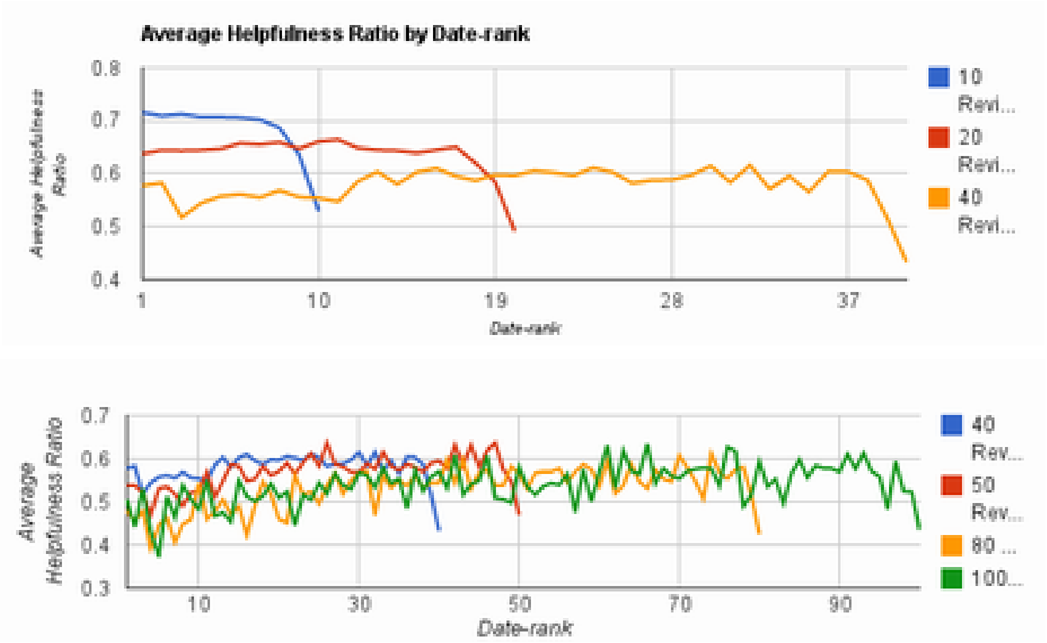


Figure 3.2: Average helpfulness ratio by date-rank. Each line represents the average deviation for a given rank for all products with the same number of total reviews.

Total reviews	10	20	40	50	80	100

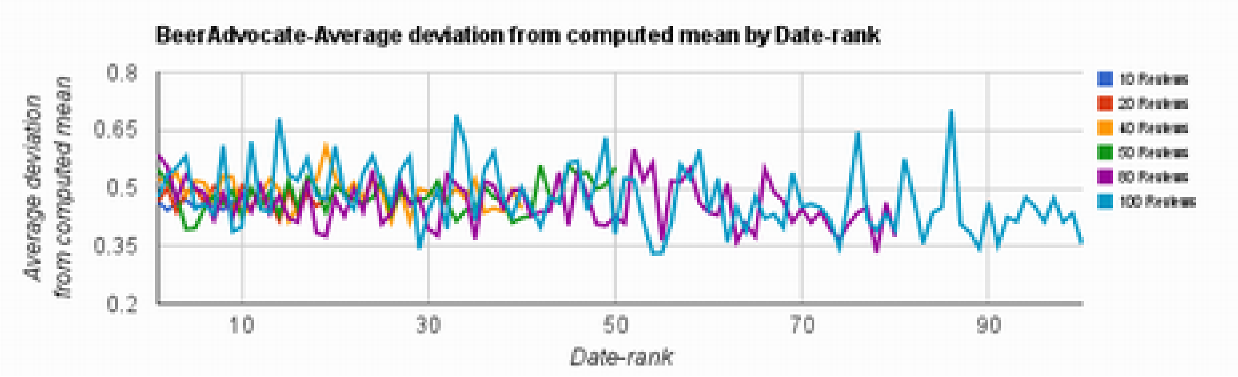
N	8668	2962	955	595	284	225
β	-0.0141	-0.0033	.0002	.0014	.0013	.0008
α	.7589	.6711	.5743	.5316	.4763	.5001

Figure 3.3: Results of simple linear regression analysis on Amazon data for average helpfulness ratio, where $y = \alpha + \beta x$. N is the total number of products included in the average deviation calculation.

Figures 3.2 and 3.3 show that for lower numbers (<20) of reviews, there is a tendency for earlier reviews to be rated more helpful. These results show that H2, the hypothesis that earlier reviews are rated more helpful, conditionally holds, with the constraint that the hypothesis holds only when there are a relatively low number of reviews; when there are a high number of reviews, there is no observable correlation between date-rank and the average helpfulness ratio.

[4.3] *Effects of the Helpfulness Mechanism: A comparative study with BeerAdvocate reviews*

Due to the observed effects of date-rank on the helpfulness ratio of a given review, an additional study is needed in order to characterize the relationship between H1 and H2 and see whether any causality can be found between the two hypotheses. Using BeerAdvocate reviews, a similar review network to Amazon, we performed similar analysis on this data set in order to see whether H1 held on BeerAdvocate reviews. In the case that a correlation between date-rank and deviation from computed average could be observed in this dataset, it can be argued that H1 holds due to an anchoring effect, rather than as a consequence of the correlation between date-rank and helpfulness ratio and the biases observed in prior work.



Total reviews	10	20	40	50	80	100
N	839	276	97	70	36	23
β	-0.0013	-0.0007	-0.0012	.0007	-0.0007	-.001
α	.7589	.6711	.5743	.5316	.4763	.5001

Figure 3.4: Results of simple linear regression analysis on Amazon data for average helpfulness ratio, where $y = \alpha + \beta x$. N is the total number of products included in the average deviation calculation. Data from which regressions are calculated is graphed above.

Key results from analysis shown in figure 3.4 indicate that no significant correlation can be observed between date-rank and average deviation from computed mean as each regression line has a near zero slope. Therefore, for the BeerAdvocate dataset, the hypothesis that earlier reviews tend to be closer to the computed average does not hold. From this, we then conclude the observation of hypothesis 1 in the Amazon dataset is caused by the observed trend described by hypothesis 2 and the helpfulness mechanism, and not by an anchoring intrinsic in earlier reviews. In other words, we observed that earlier reviews tend to be more helpful; therefore, since more helpful reviews tend to be closer to a product's final computed average, it follows that earlier reviews tend to be closer to the computed average.

[5.0] Conclusions

Conditioned on the accuracy of our results and the bounds of our "affected subset" (less than 40 reviews, encompassing ~ 91% of all reviewed products in our data set), earlier reviews are in general closer to a product's computed average compared to later reviews - an unexpected result, considering that usually large numbers are needed to create an accurate representation of some true mean. Unless earlier reviews are somehow, by fluke of "law of small numbers", consistently producing a better idea of the "true mean" for an individual product (if we are to believe that "the customer is always right", then it follows that a product's "true mean value" will be its average value as determined by the consumer), it is likely that they are, in fact, influencing the product computed average. While identifying the exact way(s) in which earlier reviews can affect the computed average of a product, the implications are clear: Amazon's helpfulness mechanism, as compared to a system without a helpfulness metric such as BeerAdvocate, causes earlier reviews to be closer to a product's computed average review, and as more helpful reviews are the most visible reviews to any customer, early reviews gain a first mover advantage, with disproportionate influence over the product's computed mean, and potentially, its market success. Future work could validate, quantify, qualify, or reject the notion of early reviews affecting a product's market success by carefully studying a new dataset composed of salesrank and top Amazon recommended products for products within our criteria, and perhaps producing a more accurate model to determine how a product "succeeds" or "fails" on Amazon, to be correlated with our results on disproportionately influential early reviews.

[6.0] References

[] Danescu-Niculescu-Mizil, Cristian and Kossinets, Gueorgi and Kleinberg, Jon and Lee, Lillian. "How Opinions are Received by Online Communities: A Case Study on Amazon.com Helpfulness Votes. 2009. http://www.cs.cornell.edu/home/kleinber/www09-helpfulness.pdf

[] Ariely, Dan and Loewenstein, George and Prelec, Drazen. "Tom Sawyer and the construction of value." 2004. http://www.cmu.edu/dietrich/sds/docs/loewenstein/TomSawyer.pdf

[] Saal, Frank E. and Downey, Ronald G. and Lahey, Mary Anne. “Rating the Ratings: Assessing the Psychometric Quality of Rating Data.” 1980. http://psycnet.apa.org/index.cfm?fa=fulltext.journal&jcode=bul&vol=88&issue=2&page=413&format=PDF