Using Properties of the Amazon Graph to Better Understand Reviews

Leon Lin
leonlin@stanford.edu

Vasant Ramachandran
vasantr@stanford.edu

Sheldon Chang
sheldonc@stanford.edu

December 11, 2011

1 Motivation

According to a 2010 paper published by PowerReviews, a e-tailing consulting company, 63% of shoppers consistently read reviews prior to making a purchase decision while 33% of that total spend at least half an hour reading reviews for a product; thus online reviews of a product have now become a strong factor in determining the sales of a particular product. Perhaps no retailer is more influenced by online ratings than Amazon.com, the world's largest online retailer, and one of the pioneering companies of online feedback scores. Amazon.com also has the world's largest collection of online reviews [1], which further emphasizes the importance of product reviews for the online retailer.

However, there has been significant controversy surrounding the accuracy and origins of many of these reviews. A 2011 article by *Daily Mail*, for instance, reports that that many manufacturers have started hiring companies to post fake reviews on Amazon.com in an attempt to increase the manufacturer's credibility while at the same time, defaming competition [6]. Trevor Pinch, a professor of sociology at Cornell, along with Web entrepreneur Filip Kesler published a paper further strengthening the assumption that there is personal bias involved in many of the reviews, by uncovering the fact that around 85% of Amazon's top 1000 reviewers receive free products from publishers, agents, authors, and manufacturers [2].

The helpfulness rating of the reviews themselves have also been found to be bias. [3] demonstrates that user helpfulness ratings oftentimes do not directly reflect the content quality of the review itself. [4] elaborates on this notion by illustrating the fact that users rate the helpfulness of a review based upon their own personal preference towards the item.

This project aims to accomplish two tasks: to be able to predict future Amazon ratings given current rating data, and to adjust the current star rating scheme to better reflect the true quality of the item. Predicting the future rating of a product is valuable because it gives both manufacturers and customers the opportunity to know the consensus perception of the product at an earlier time. Being able to establish when the future rating can be predicted with a high degree of confidence is also desired since many customers will not purchase a product until they are confident that the consensus view of the product is a favorable one. In fact, [1] found that 72% of consumers find user product reviews to be a "very/extremely" important factor when it comes to selecting and purchasing a product.

Because a vast majority of users rely upon Amazon.com reviews to determine the quality of a product, attempting to produce more accurate ratings by accounting for the biases mentioned above is also an important task. The star rating likely represents the first opportunity most users have to pass judgement on an item they find when browsing Amazon, and the star rating also factors into other components which can influence the sale of a product, such as when that product appears in the search results.

2 Data

2.1 Amazon Data

While there exists multiple sources from prior studies which provide data about Amazon [12] [13], none offered the depth of information we needed to reconstruct a graph of users and their reviews for products. To obtain the novel information we wanted, we used a combination of the Amazon Associates API, as suggested by [4] in conjunction with a customer scrapper we built. A custom scrapper was necessary because in 2009, Amazon no longer made it possible to fetch review or reviewer information from their API. Amazon also implemented a cap on the information which could be retrieved within a given time frame, both through the use of their API and through a web crawler that scrapped the HTML from their page. Thus, our implementation required building a program that periodically fetches information both through the API and Amazon webpages.

We ran the scrapper for about a week and ended up retrieving 72,859 unique product reviews, reviewed by 63,962 unique reviewers. For each review, we retained the ASIN, a unique identifier of an item, as well as the nearest parent browse node of the item (essentially the closest thing to the category of an item), the rating given by the review, the time the review was posted, and number of users who found the review helpful, the number of people who found the review unhelpful and the reviewer ID for the item. Reviews are uniquely identified using the reviewer ID and item ASIN. Each reviewer contained the reviewer's ID, the total number of users who have found the user helpful and unhelpful, and all the badgers obtained by the reviewer.

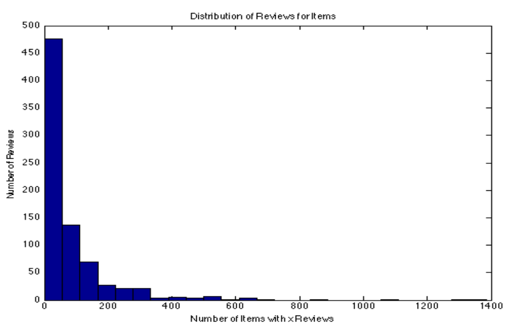

Figure 1: Histogram of # Reviews for each Item

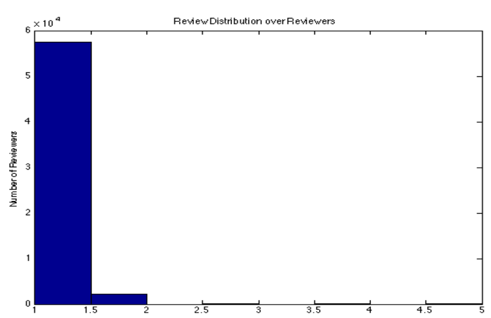

Figure 2: Histogram of # Reviews Written by Each Reviewer

2.2 Alatest Data

We turned to the review aggregation site alatest.com for expert reviews. Alatest collects both user and expert reviews from around the web for a wide range of consumer products, such as computers, televisions, cameras, video games, and other types of electronics. Expert reviews are drawn from CNet, PCWorld, and a variety of other electronics review sites and blogs, while consumer reviews come from sites like Amazon, BestBuy and CNet. For each product, Alatest displays snippets of written reviews and compiles average user and expert ratings (both out of ve stars), as well as a calculated ALA Score (out of 100). This ALA Score is both weighted and normalized for all products in the category, and accounts for various factors such as the number, strength, and source quality of the reviews.

The data was scraped by using the Python web crawling and scraping library Scrapy. The index for the site was fed to the crawler, which iterated over all product categories and gathered the top 500 products in each category and discarded any products that did not have one of the following: a user average, an expert average, or an ALA score (not all products have these fields due to lack of data). This left us with a database of 16,000 products. In order to ensure that the average expert review was statistically significant, we further limited this dataset to include only products with at least 20 expert ratings, which cut down the number of products to 3,733.

3 Create New Metrics

While quantifying the impact of any one review on the visibility, reputation, or sales performance of a particular product is not a straightforward problem, some metric of helpfulness to the average user on Amazon is necessary to assess the value or validity of any review in relation to the ideal. Amazons feedback system is relatively simple, with a binary classification system for helpfulness. Though, as with many feedback mechanisms, Amazon is not immune to the user propensity to rate an unhelpful review helpful, the proportion of helpful to (unhelpful + helpful) votes for a particular review is a useful starting point to determine the value of a review to a potential buyer.

However, this proportion does not capture the law of large numbers as implicit in the feedback data. It is clear that the extreme ends of the helpfulness proportion distribution belong to those whose reviews have gotten very little feedback, which increases the likelihood of no helpful votes, or no unhelpful votes. However, as the number of feedback votes increases, the review will not stay at the extremities of the distribution. Thus, we chose to use two different metrics for both the review feedback and the aggregate feedback for a given reviewer, where one metric simply recorded the proportion of helpful votes to all votes, while the other multiplied this proportion by the fraction of maximum possible votes over all reviews(or reviewers) that this review(reviewer) received. Drawing our hypotheses from literature on this subject, we conjectured that reviews which tracked the mean review for a given product would obtain higher helpfulness proportions and that negative reviews would be somewhat more immune than positive reviews from this phenomenon because of the perception that they were brilliant but cruel [4]. We also hypothesized that a small proportion of reviewers would receive high exposure and a very large number of feedback votes for their reviews, following the Zipfian(power law) pattern often seen in social network degree distributions.

As an analysis of graphs shown below implies, the relative scarcity of feedback for the vast majority of our reviewers skews the helpfulness proportion distribution towards the extreme ends (0 and 1), particularly for specific reviews (Figure 2). While aggregates for reviewers (Figure 3) are not quite as skewed and smoothen out the noticeably discrete histogram of (Figure 2), this effect is still noticeable. It is clear that most reviewers receive more helpful than non-helpful votes, as predicted before, while the predicted Zipfian distribution for votes received is apparent in the highly skewed normalized to the maximum version (Figure 4) of the helpfulness distribution, where each proportion is essentially weighted by the number of votes it received as a fraction of the maximum observed votes for any given review. To test the conformity and brilliant-but-cruel hypotheses that might mar our use of the helpfulness proportion as a predictor of the ideal review, we found a correlation (r^2 of approximately 0.68) between the inverse of the deviation from the mean and the helpfulness proportion, suggesting that the conformity hypothesis may be significant in our data set. On the other hand, however, more-negative-than-average reviews only saw a 0.28 r^2, suggesting that negativity may not validate a review to users as brilliant.

3.1 Bias and the Ideal Review

Quite separate from a metric of utility are the questions of bias and data skew in the reviews themselves: some excellent user-ratings may still reflect a user bias towards very negative ratings in a particular product area, or a proclivity towards companies returning favorable reviews with favorable reviews. When considering the reviews, [4] suggest that reviews tend to be drawn from two separate distributions, one for those that view the product favorably, and another for those that do not. Empirically, they show that as the variance for product ratings increase, the ratings distribution becomes bimodal. They develop a simple model that relates these two distributions with two parameters: balance (percentage of positive reviews p and negative reviews (1-p) and controversy–how far apart the means of the distributions are) They demonstrate that this model reflects the actual ratings distributions as the controversy parameter is increased. The result is a distribution that is at first unimodal, and then bimodal as the controversy increases.

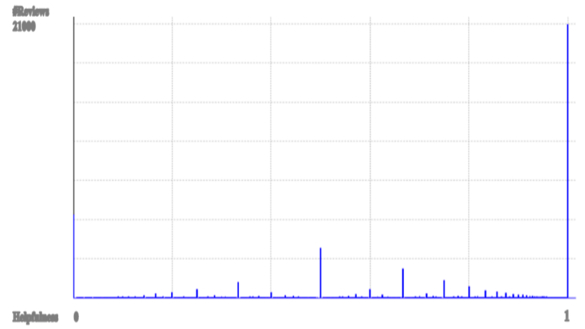

Figure 3: Distribution of Helpfulness Proportion Amongst 72,859 Reviews

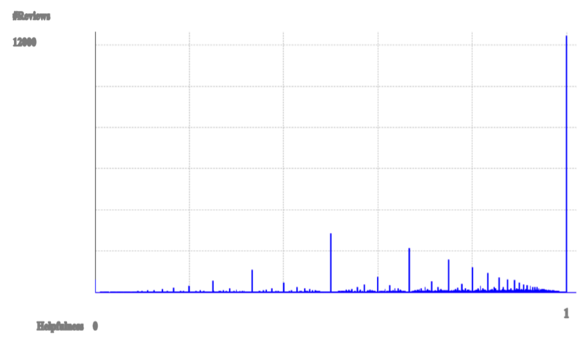

Figure 4: Aggregate Helpfulness Proportion Distribution for Reviewers

Our study of systematic reviewer bias showed that aggregate reviewer bias closely mirrors the deviation of individual reviews from the average for a given review. This deviation appears to follow a standard Gaussian distribution, though there is undoubtedly some skew, suggesting that a few extremely negative reviews pull down the average review for any given product, and that the disproportionate number of positive reviews may mean that it is simply impossible under Amazons five-point system to give an excessively positive review that is far removed from the mean. The five point system may also represent

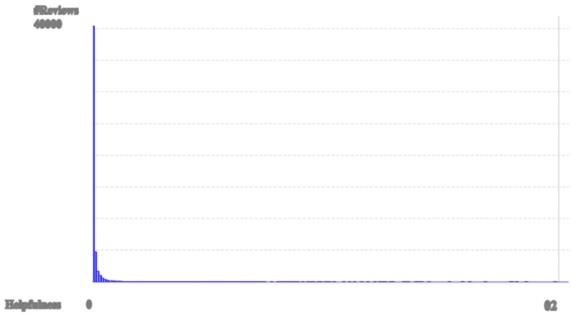

Figure 5: Helpfulness Proportion Normalized to the Maximum Number of "Helpful" votes received

a low level of controversy, which would explain why the distribution appears unimodal. Below is the distribution of aggregate average reviewer star deviation from the mean star rating(over all reviews written by the reviewer) for each review.

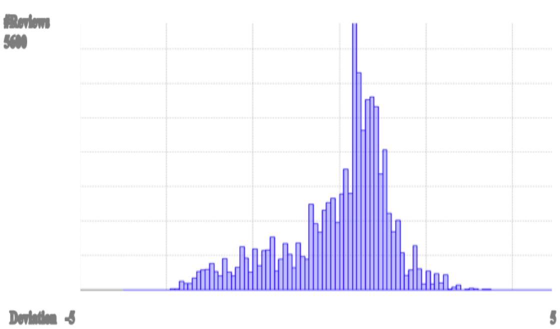

Figure 6

3.2 Reviewer Expertise

Perhaps the most difficult question to answer with relatively sparse data about reviewer predispositions and relatively few prolific reviewers was the question of expertise about a certain category of product, which would enhance the value of any review written by an expert. We wished to study this feature of a review in isolation from the perception of helpfulness by feedback voters. However, as we only were able to obtain information about the prior reviews of any given reviewer and the category of product under which those reviews fell, we hypothesized that prior interest in reviewing a particular category of product implied some form of expertise through experience.

In order to identify users particular qualification to rate a product, we realized that strong preference for products of the type being reviewed accentuates interest and ability to write a good review, as well as trust from users of the product rating system to make purchase decisions [8]. We decided to take this expertise into account using a collaborative filtering approach.

Collaborative filtering attempts to make predictions about users preference or interest from the behavior of other users similar to the user of interest. In the case of Amazon, we defined behavior affirming interest as simply reviewing the item, and we assigned similarity scores between reviewers based on numeric overlap in products reviewed (normalized by total products reviewed), as well as similarity in rating averages. Thus, with our system, a user reviewing an item which had not been reviewed by users similar to him in rating average and other products reviewed would not be seen as an expert, but rather as an amateur. Each review receives a CF score that establishes the strength of the relationship between reviewer and item-reviewed. Our relatively sparse data set (in terms of number of reviews per reviewer) allowed this relatively time-intensive computation.

We used the Pearson correlation coefficient collaborative filtering routines built by Dr. Guy Lebanon at Georgia Tech as a toolkit in MATLAB [15]. Scores were normalized to a minimum value of 1 and a maximum value of 5 (to correspond to the Amazon star system) which would indicate that all reviewers who reviewed the item of interest received perfect similarity scores with the reviewer of interest. Below is the distribution of CF expertise scores over all reviews.

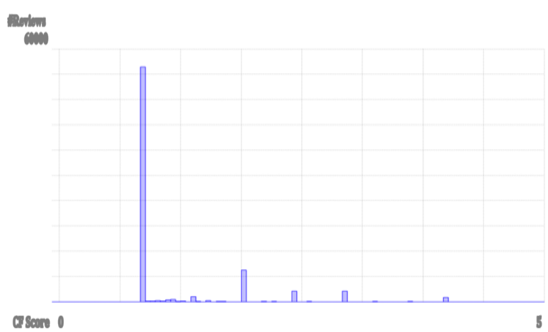

Figure 7

Again, our high reviewer-review ratio makes it difficult to ascertain the expertise of most of our reviewers, who may receive low scores due to sparse data about the review preferences of reviewers similar to them. However, the expertise level, even when determined from this rudimentary approach, clearly shows the power-law distribution of mavens on a given topic.

3.3 The Reviewer Graph

Modeling the reviewer-item network as a graph allows for the investigation of additional predictors of ideal review-authorship related to the behavior of reviewers as components of a social network. In this network, edges connect reviewers who have reviewed the same item: there is a direct relationship between the number of items reviewed and the degree of the reviewer.

We hypothesize that the degree distribution of the reviewer set gives some insight into the validity and approach to authorship of his or her reviews: a high degree may simply mean that the reviewer chooses to write only upon items about which there is already extensive insight or information provided by other users. On the other hand, a low degree may indicate an interest in reviewing relatively invisible items with low review authorship: niche products or maven-focused items. An additional metric of interest is the reviewers betweenness centrality or propensity to connect otherwise disparate components of the graph in a shortest path. Our calculation of this metric utilized the Brandes algorithm for its efficiency [16]. In this context, high betweenness-centrality may indicate eclectic interests or expertise in a wide range of topics. The distribution of degree among reviewers (Figure 7) is surprisingly even over a wide range of values, while the distribution of betweenness centrality(Figure 8),with relative uniformity followed by a sharp drop towards the tail, suggests the presence of clusters of reviewers with particular interests and proclivities, with relatively few reviewers able to bridge the gaps between the clusters.

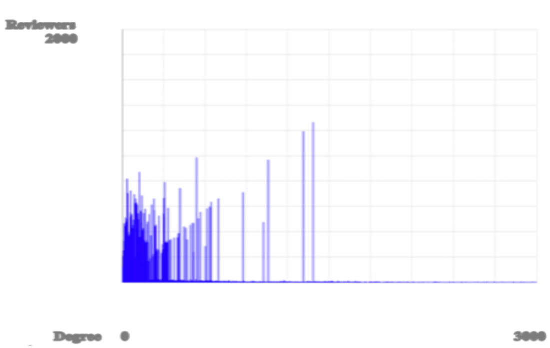

Figure 8

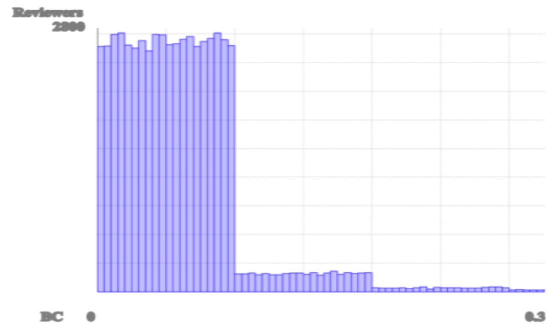

Figure 9

4 Predicting Future Feedback Scores

4.1 Methodology

Because predicting future feedback scores based upon prior feedback scores is only applicable to products with a moderate number of reviews, we filtered our data such that only products with at least 100 reviews were retained. Having retained 191 product reviews, for each of those product reviews, we sorted those reviews by date and discarded any reviews that occurred after the first 100 reviews.

To obtain the "true" average rating for a product, we averaged the scores of the first 100 ratings. We then obtained a baseline for our model by observing how the average feedback score of the first m ratings compared to the true average feedback score we computed. We attempted to improve this feedback score by calculating modified ratings as a weight sum of many of the features highlighted in the previous section, including the helpfulness and of the review, the helpfulness of the reviewer, how bias the user is, the expertise of the user, the original star rating given by the review, as well as the betweenness, degree, and clustering metrics previously discussed. We note that the review bias was not used and expert bias was modified as these metrics, in their original forms, take into account knowledge of the true average rating of a product.

In total, we took into account 10 parameters. We solved for the weights of each of these parameters in computing a modified feedback score by minimizing the following objective function:

$$(y - Ax)^T (y - Ax) \tag{1}$$

where A is a $n \times p$ matrix of features and x is a $p \times 1$ vector of feature weights. We note that n is the number of items in our data, while p is the number of features used, and so $n = 191$ and $p = 10$ in our data set. The vector y is a $n \times 1$ vector of true means. In other words, for a particular training example, k, the value y_k corresponds to the true mean of the item with a review which has features, $A_{k,i}$ for $i = 1, \ldots, p$. We solved for the value of x by calculating the pseudo-inverse of A. For rating i of an item a, which we will denote a_i, we attempted to calculate a modified rating, a'_i using the following equation:

$$a'_i = A_{(a,i)} x \tag{2}$$

Where

We also attempted to find the bias of a rating with respect to when the ranking occurred. Let t represent the time at which a review occurred (so if $t = 5$ for a review, that review was the 5th earliest review for an item). We now want to be able to find a function $f(t)$, which allows us to account for any potential bias in reviewing an item at a certain time. $f(t)$ was calculated by observing the average amount of deviation of ratings at each time step from the true average rating of a product. $f(t)$ was then used to offset any potential bias by either subtracting the value $f(k)$ from any rating that occurred at time k.

These modifications to the rating of items were then tested using 3-fold cross validation, where the training set was used to calculate both x in (1) and $f(t)$, and the parameters were applied to a held out test set.

4.2 Results

The following graph illustrates how the current average rating (where the current average is simply the average of all the reviews that occur at time $t \leq x$) of a product deviates from the true average given different rating schemes. The baseline uses the original ratings, while the "with $f(x)$" scheme uses weights which offset bias from reviews occurring at certain times. The "modified rating" scheme uses the modified ratings as described by (2), and the "$f(x)$ and modified rating" scheme adjusts both for time dependent bias and uses the modified ratings. Specifically, this scheme uses the time independent rating as a feature rather than the time dependent scheme.

It should first be noted that while the baseline and scheme which accounts for time dependent bias converges to the true mean after 100 time steps, the schemes which use modified rating systems do not. This is because the true mean rating remains the same regardless of whether the original ratings are modified or not. These deviations were calculated from averaging the deviations from each test set in each of the cross validation steps. As can be seen,

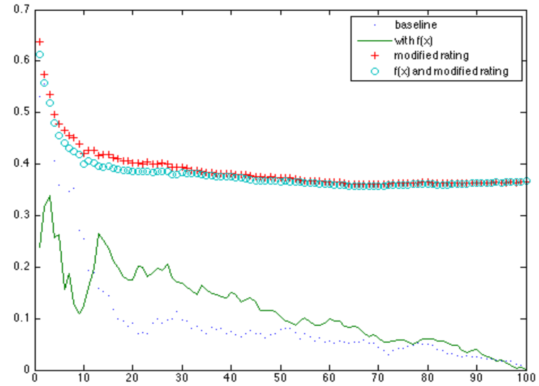

Figure 10: Deviation of Various Rating Schemes

neither scheme using the modified ratings were able to outperform the original rating scheme.

The most glaring trend in observing the baseline score is the fact that early product reviews have a much higher chance of rating a product favorably than later produce reviews. This trend is illustrated in the graph below,

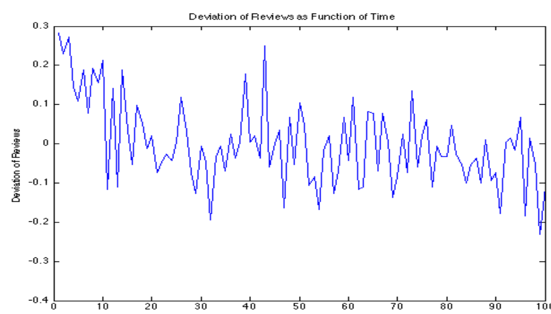

Figure 11: Deviation from Average Star Rating

Note that unlike Figure 9, Figure 10 does not take the absolute values of the deviations, thus allowing negative and positive deviations to be noticed. More precisely, the Spearman correlation between the deviation from the mean at each time step and t was found to be -0.3928, showing a rather drastic relationship between the ratings given by a review and the time at which it was given. Using the training set to compute the average deviation from the true mean at each time step, we were able to compute a rough estimate for how bias a rating is for a given time step, and use that bias to smooth out ratings in the test set. As can be seen in Figure 9, the time independent rating scheme actually outperforms the baseline in predicting the true mean for $t < 12$.

To test the real world applicability of our modified rating scheme, we created a new test that takes in two parameters, α and β. The test works by examining the first 99 reviews, calculating the average of those reviews, and seeing how far that average deviates from the true mean. The test continues removing reviews until α percent of the time steps tested have averages which deviation by more than β percent of the true mean from the true mean. The test then outputs the time step at which it stopped for each item, and averages the stopping time step at for all items given the set of parameters. Lower scores are considered better on this test, as an intuitive, though not completely accurate, interpretation of the test is, how many reviews do you need before you can say, with accuracy α that your estimate of the true average rating for a product will be within β percent of the true value. Note that this test could not be performed with the other schemes which modified the ratings of each review as the average value of those modified rankings did not converge to the true mean.

Parameters	**Baseline**	**Time Ind.**
$\alpha = 0.75$, $\beta = 0.25$	1.691	1.345
$\alpha = 0.85$, $\beta = 0.15$	4.680	4.209
$\alpha = 0.90$, $\beta = 0.10$	10.33	10.11
$\alpha = 0.95$, $\beta = 0.05$	34.48	34.69
$\alpha = 0.99$, $\beta = 0.01$	84.84	86.88

As we can see in the results, the time-independent rating scheme outperforms the baseline scheme when only wanting a fairly rough idea of how how well the product will eventually be rated, however, as we require more precision and gain more reviews, the baseline scheme outperforms the time-independent one.

Finally, we wanted to also list below the weight given to each feature averaged over each cross validation step for our calculation of the modified weighting scheme. Both modified weighting schemes gave similar weight values, so we list only the weights for the time independent case. Note that the weights are directly comparable since each feature was standardized to have a mean of 0 and variance of 1.

Feature	**Weight**
Original Rating	155.7351
Review Helpfulness Count	-4.0034
Review Unhelpfulness Count	25.2860
Review Helpfulness Ratio	10.4437
Reviewer Helpfulness Ratio	-1.5782
Reviewer Bias	-45.0138
Reviewer Expertise	-7.8396
Reviewer Clustering	21.0553
Reviewer Betweenness	1.8424
Reviewer Degree	-26.4712

4.3 Discussion

Unfortunately, the metrics which we calculated did very little to aid in the prediction task. This is somewhat expected as the true means are calculated directly from the original rating scheme, and thus it would be difficult to predict this mean with modified ratings. Nonetheless, we can still gain some insight by looking at how the weights were distributed. As expected, by far the strongest weight is the original rating of the review. From here, we can see that many of our weights make intuitive sense. For instance, reviews which tend to be considered the least helpful are those usually give a product a highly negative rating Thus it makes sense for the unhelpfulness of a review to have a positive weight since unhelpful ratings tend to rate a product below the true mean of the product and thus needs a positive weight to have a modified rating closer to the mean. Similarly, we see that the second strongest weight belonged to reviewer bias, and that bias reviewers tend to review an item higher than user. This finding is particularly interesting as it corroborates one of the findings discovered through a different means: that early reviewers have a bias for reviewing items highly. Another interesting fact to note is that our findings directly contradict the “Brilliant but cruel” assumption mentioned by [4], along only weakly, as reviewers with greater expertise tended to rate an item more highly.

The finding that earlier reviewers tend to give more positive reviews is an interesting finding. Intuitively, this can be explained by the fact that early adapters are buying items before they even read reviews or know the public perception of an item, and thus have a natural tendency to view the item favorably. Perhaps most importantly, our test shows that accounting for this early adopter bias tends to allow for better prediction of the true product mean from a smaller subset of data. Our method consistently outperforms the baseline when given only a small data set, a very valuable property considering the vast majority of items on Amazon have very few reviews.

5 Predicting Expert Ratings

5.1 Motivation and Methods

We explored the use of these calculated metrics in the task of coming up with a more accurate way of determining the "true rating" for a product. On Amazon, a simple average over all user ratings is displayed for each product, but this rating is susceptible to fraud, as every rating is weighted equally – as noted in the introduction, "review stacking" on Amazon is a problem as authors or others with a vested interest in a product will often produce dummy accounts and leave false ratings and reviews [6]. By taking into account factors such as the expertise, bias, and helpfulness of the reviewer, we hoped to introduce weighting scheme that produced a final rating that was more indicative of the true rating for a product. As the gold standard in this task, we used the expert ratings average drawn from alatest.com.

The average user rating was used as our baseline predictor for the expert rating. For our learning algorithm, we chose to use both linear regression and decision trees with 10-fold cross validation.

5.2 Discussion and Results

The data was first modeled using simple linear regression. Each training example was a feature vector that contained factors including user rating, number of helpful and unhelpful votes, reviewer bias and expertise. The average user rating was used as a baseline, and when it was the sole feature, had a correlation coefficient of 0.2809 and a root mean squared error of 0.3458, over 72,600 reviews drawn from a variety of categories. Incorporating the rest of the features yielded a slight improvement, with a correlation coefficient of 0.3133 and RMS error of 0.3422. Using decision trees produced no improvement over the baseline, with both yielding a RMS of 0.3381 and a correlation coefficient of 0.3454. The most heavily weighted factor in the linear regression, and the feature that produced the most information gain in the decision trees, was the original baseline metric of the average review rating (Figure 11).

Our best results were obtained by pruning the labeled examples to include reviews with only 100 or more helpfulness votes. This yielded 617 reviews. While sparsity was an initial concern, in that some products would only have one review, this turned out not to necessarily be the case, as products that had

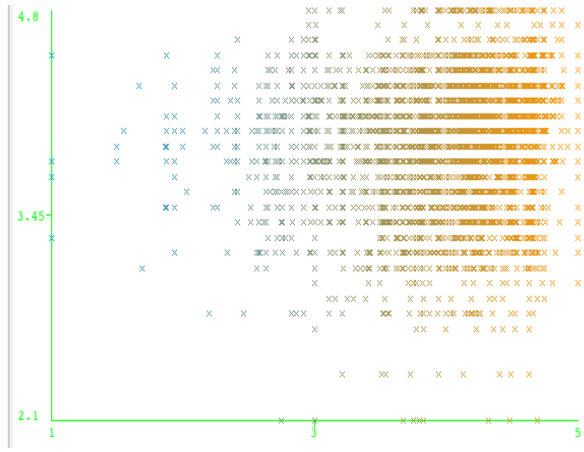

Figure 12: The Most Highly Correlated Variable (Average User Rating) Versus Expert Rating

reviews that earned more than 100 helpfulness votes tended to be very popular items that had several such reviews. Linear regression using this pruned data set yielded a baseline correlation coefficient of 0.2018, but incorporating the other factors boosted this to 0.3811. Favorable results were also obtained by selecting examples with an expertise rating of greater than or equal to 2. There were 13,934 such reviews, and incorporating all these metrics beat the baseline by a correlation coefficient of 0.18 versus 0.07.

While these results are not outstanding, they are nonetheless statistically significant. The metrics that had the best impact on predicting expert ratings were helpfulness votes and expertise ratings, while the other calculated metrics had little to no impact. From an intuitive standpoint, it makes sense that these metrics would be more accurate in predicting the expert rating, as both the expertise score and number of helpfulness votes are themselves indicators that the reviewer is more of an expert than ordinary.

[4] proposed a simple mathematical model that hypothesized that reviews were drawn from two underlying distributions, one for positive and another for negative reactions to the product. It would be interesting to incorporate this framework into our expert review predictions, as parameters in this model include the balance and controversy between positive and negative reviews. These parameters could also be incorporated as features, to see perhaps if high controversy, for example, is associated with more negative expert ratings.

References

[1] P. Marsden. "PowerReviews 2010 Social Shopping Study." *Social Commerce Today.* PowerReviews, 14 May 2010. Web.

28 Oct. 2011. http://socialcommercetoday.com/2010-social-shopping-study-top-line-results.

[2] T. Pinch, F. Kesler. *How Aunt Ammy Gets Her Free Lunch: A Study on the Top Thousand Customer Reviews at Amazon.com.* N,p., 2010. Web.

[3] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou Low-quality product review detection in opinion summarization. In *Proc. EMNLP-CoNLL*, pages 334342, 2007. Poster paper.

[4] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, L. Lee. How Opinions are Received by Online Communities: A Case Study on Amazon.com Helpfulness Votes. In *WWW*, 2009.

[5] Nan Hu, Jie Zhang, and Paul A. Pavlou. 2009. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* 52, 10 (October 2009), 144-147. DOI=10.1145/1562764.1562800 http://doi.acm.org/10.1145/1562764.1562800

[6] "Five stars for effort: Amazon and TripAdvisor at centre of scandal as companies post fake reviews." *Daily Mail*. N.p., 2 June 2011. Web. 28 Oct. 2011. http://www.dailymail.co.uk/news/article-1393412/Amazon-TripAdvisor-centre-scandal-companies-post-fake-reviews.html?ito=feeds-newsxml.

[7] S. David and T. J. Pinch. Six degrees of reputation: The use and abuse of online review and recommendation systems.

[8] P. Massa and P. Avesani. Trust-aware collaborative ltering for recommender systems. *LNCS*, 3290:492508, 2004

[9] Patricia Victor, Chris Cornelis, Ankur Teredesai, Martine De Cock: Whom should I trust?: the impact of key figures on cold start recommendations. *SAC* 2008: 2014-2018

[10] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2001). "8.5 The EM algorithm". *The Elements of Statistical Learning*. New York: Springer. pp. 236243.

[11] R. M. Bell and Y. Koren. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In *ICDM-07, 7th IEEE Int.Conf. on Data Mining*, pages 43-52, Omaha, Nebraska, USA, 2007.

[12] Nitin Jindal and Bing Liu. "Opinion Spam and Analysis." *Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008)*, Feb 11-12, 2008, Stanford University, Stanford, California, USA.

[13] John Blitzer, Mark Dredze, Fernando Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. *Association of Computational Linguistics (ACL)*, 2007

[14] alaTest. N.p. Web. 28 Oct. 2011. www.alaTest.com

[15] Prem Melville and Vikas Sindhwani. In *Encyclopedia of Machine Learning*, Claude Sammut and Geoffrey Webb (Eds), Springer, 2010.

[16] Ulrik Brandes: A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology* 25(2):163-177, 2001.