Para 'normal' Activity: On the Distribution of Average Ratings

Nilesh Dalvi
Facebook
Menlo Park, CA
nilesh.dalvi@gmail.com

Ravi Kumar and **Bo Pang**
Google
Mountain View, CA
{*ravi.k53, bopang42*}*@gmail.com*

Abstract

In this paper we study the distribution of average user rating of entities in three different domains: restaurants, movies, and products. We find that the distribution is heavily skewed, closely resembling a log-normal in all the cases. In contrast, the distribution of average critic rating is much closer to a normal distribution. We propose user *selection bias* as the underlying behavioral phenomenon causing this disparity in the two distributions. We show that selection bias can indeed lead to a skew in the distribution of user ratings even when we assume the quality of entities are normally distributed. Finally, we apply these insights to the problem of predicting the overall rating of an entity given its few initial ratings, and obtain a simple method that outperforms strong baselines.

Introduction

When it comes to choosing among several options, the web, over the last decade, has tacitly conditioned most of us to seek online ratings/reviews of the options. Imagine the last time someone purchased a product online without studying its reviews or went to a restaurant ignoring its ratings or watched a movie without hearing the opinions about it. And, imagine the countless human hours expended in perusing the reviews from multiple websites and experts in order to make a choice. (Ironically, in many cases, the purchase price pales in comparison to the value and the amount of time invested in making the purchase!)

The choice behavior of users is becoming so dictated by ratings that sorting by average rating is presented as the default option on most e-commerce sites. Average ratings are equally important from the point of view of businesses/products. A study by Luca (2011) showed that when an independent restaurant's average rating increases by one star on Yelp, its revenue goes up by 5%–9% (see also http://www.mainstreet.com/article/lifestyle/food-drink/real-value-yelp-review); more strikingly, the actual review text only plays a secondary role and the average rating takes on the primary role. Anderson and Magruder (2012) showed that an extra half-star causes restaurants to sell out 49% more frequently. Another study by Reinstein and Snyderz (2005), Moon, Bergey, and Iacobucci (2010), and several others have showed, adjusting for various factors, that positive ratings on movies have a positive impact on the revenue. Given these, the role of average rating of an entity, be it a movie, product, or restaurant, becomes an important quantity to study from both economic and human behavioral points of view. What are the characteristics of average ratings and how do they arise?

The starting point of our work is an analysis of average user ratings from three distinct domains, namely, restaurants, movies, and products. We study the distribution of average user ratings in each of these domains. We show that the distribution in each case is heavily right skewed, with a surprisingly close fit to a reflected log-normal distribution; to the best of our knowledge, this observation is new. Furthermore, for the movie domain, we compare the distribution of average critic ratings and average user ratings, and find the former to have a much closer fit to a normal distribution. This suggests that the log-normal shape cannot be explained solely by the quality of entities; characteristics of rater behavior clearly play a role. We then proceed to study the user behavioral process that can lead to the observed skew.

We propose *selection bias* as the main reason for obtaining a log-normal user rating distribution. The theory behind selection bias is that users select entities that they expect to like and hence rate them positively. We formulate selection bias by using a simple model: a user applies a selection function that determines how likely he/she is to select the entity. This model is closely related to the econometric correction model of Heckmann (for which he won the Economics Nobel prize in 2000). We show that a simple threshold function for selection is enough to lead to a log-normal user rating distribution, when the underlying quality of the entities follows a normal distribution; our observation about the critic rating (which is a proxy for the underlying quality) distribution lends credence to the latter assumption.

We then apply these insights to the following important problem: given the first few ratings of an entity, predict its final average rating. For this problem, we obtain a simple algorithm that uses our observations, and show that this algorithm performs better than strong baselines such as the average or the weighted average.

Related Work

The related work falls into the following four categories: work on selection bias, the economic impact of ratings, the distribution and temporal characteristics of ratings, and the problem of predicting popularity from early user feedback.

As mentioned in the Introduction, the study of selection bias is by now classical; see the Wikipedia article (http://en.wikipedia.org/wiki/Selection_bias). Heckman (1979) developed a number of statistical methods to correct for selection bias; our basic model is motivated by his pioneering work, though the end applications in his case are related to regression-type problems. A nice survey of the model, along with extensions to other settings, can be found in (Dubin and Rivers 1989). In online settings, Kramer (2007) observed that self-selection leads to better-than-average ratings in reputation systems. His focus was only on the mean of the average rating distribution and on a mechanism to offset the bias in the mean; we, on the other hand, are interested in studying the entire distribution of the average rating. Li and Hitt (2008) also noted that self-selection biases, especially of the early adopters, may negatively affect long-term consumer purchase behavior. They developed a model of self-selection by the early adopters and analyze its impact on future sales, whereas we treat self-selection as a process that happens for each user rating (not just of the early adopters). Berinsky (1999) argued that selection bias might cause differences between collective public sentiment and aggregated public opinion. The former can be thought of as the true quality of an entity and the latter its observed rating; however, he did not formulate any model to capture this. Sikora and Chauhan (2011) presented a Kalman-filtering based technique to estimate the sequential bias in online reviews. Their model is different from ours in that they assumed that the current review is biased by the previous review. Ma and Kim (2011) addressed the self-selection bias, but in the context of users providing multiple reviews.

The question of bias in ratings has also been addressed by the Recommender Systems community. For a theoretical economics view on this problem and how to use market mechanisms to solve them, see (Christopher, Resnick, and Zeckhauser 1999). An empirical study of non-random user rating behavior (influenced by their opinion about the entity) was done by Marlin et al. (2007) and the effect of non-random missing data (due to selection or rating bias) on rating-based recommender systems was studied by Marlin and Zemel (2009). None of these works attempted to model the selection bias *per se*.

Even before the era of large-scale online reviews, researchers had already studied the relationship between market performance and the role of critics (Eliashberg and Shugan 1997). There is a rich literature discussing the economic impact of online reviews; see (Pang and Lee 2008) for a survey on early work. As mentioned in the Introduction, Anderson and Magruder (2012) showed that restaurants that just barely get four stars sell out about 19% more frequently than restaurants that almost get four stars. Their findings contradicted earlier reports by Hu, Liu, and Zhang (2008), who found that the impact of online reviews on sales diminishes over time. A U-shaped relationship between the average propensity to review a movie and box office revenues was observed in (Dellarocas, Gao, and Narayan 2010): moviegoers appear to be more likely to contribute reviews for very obscure movies but also for very high-grossing movies.

The J-shaped distribution of raw ratings was documented by Hu, Zhang, and Pavlou (2009), who proposed purchasing bias (similar to choice-supportive bias) and under-reporting bias (only the extreme raters expressing a view) as reasons for the shape; see also (Hu, Pavlou, and Zhang 2006). Unlike our work, they did not work with the average rating of the product, but instead with the raw ratings. Similar observations about the skew in raw ratings were made earlier by Chevalier and Mayzlin (2006) and Kadet (2007). The distribution patterns of ratings have been used to study opinion spam and detect deceptive reviews (Jindal and Liu 2008; Feng et al. 2012); these merely used the empirical form and did not in particular delve into the analytical details. Wanderer (1970) studied the movie review patterns of critics and users and used that to refute the snobbism of critics. Ott, Cardie, and Hancock (2012) proposed a model of reviews based on economic signaling theory, in which consumer reviews diminish the inherent information asymmetry between consumers and producers, by acting as a signal to a product's true, unknown quality. The temporal dynamics of ratings has been studied in various contexts. Godes and Silva (2012) observed that the average rating for books decreases over time and proposed models to explain this; we observe a similar behavior with window-averaged restaurant ratings and in addition, we observe that the variance increases. Wu and Huberman (2010) observed that later opinions tend to show a big difference from earlier opinions, which moderates the average opinion to the less extreme.

Previous work has also addressed the *early prediction problem*, where popularity of social media content was predicted based on early user feedback. The specific definition of popularity could differ, however. The formulation closest to ours is the work of Yin et al. (2012). They quantified popularity as the eventual average rating an entity would receive, and considered the problem of ranking entities by their predicted popularity from the early (binary) votes they received. They proposed a model to infer the latent types of the users and used that to do a careful weighted average of the early votes. However, their focus was to find the *top* entities that would eventually become highly rated. In contrast, in our formulation of the early prediction problem, we focus on accurately predicting the final average ratings for all entities. Other previous work has quantified popularity by the magnitude of user attention such as the number of votes or views an entity received. Lerman and Hogg (2010) addressed the problem of predicting popularity based on early user reactions to new content and developed stochastic models of user behavior on social media sites for this task. Szabo and Huberman (2010) studied early prediction of popularity for Digg stories and YouTube videos and found strong correlations between popularities at early and later times. More recently, Pinto, Almeida, and Gonçalves (2013) studied the same problem over YouTube data, and improved the prediction accuracy by accounting for the temporal dynamics

Dataset	# of ratings	# of users	# of entities
YELP	1,817,018	256,343	21,731
AMAZON	3,400,317	1,402,563	109,645

Table 1: YELP and AMAZON rating data.

in the observed early viewing patterns. None of these dealt with predicting the future average ratings or modeling the role of selection bias.

Data

Our main datasets are rating data from two domains: local businesses (YELP) and consumer products (AMAZON). Entities in these domains are consumed at relatively steady speed, compared to, say, the movie domain, where each entity is mostly consumed during a short and concentrated period of time. We wanted to focus on these type of domains since they allow studies of temporal effects. Information on the size of these datasets is summarized in Table 1.

YELP We obtained a random sample of 1.8 million ratings from reviews authored by 256,343 users for 21,731 entities on yelp.com. Of these, 14,926 entities received at least 10 ratings and 39,005 users have provided at least 10 ratings. Each rating is an integer between one and five, and is accompanied by a timestamp.

AMAZON We extracted a subset of ratings from the *Amazon Product Review Data*[1] as follows. We picked those products with at least five ratings, and extracted over 3 million ratings associated with these 109,645 products. These ratings were provided by over 1 million users, 16,387 users provided at least 20 ratings. 46,845 products received at least 20 ratings. Each rating is between one and five stars, in increments of half star.

IMDB In addition, we extracted average user ratings and critic ratings for movies from IMDb. IMDb provides an average score of user ratings once a movie has received at least five ratings. Note that not all ratings are accompanied by textual reviews[2], and we do not have information on either the rater or the timestamp for each individual rating. In addition, a subset of the movies also has a *Metascore*, which is the average of critic ratings collected by metacritic.com. For the movie shown below, the average user rating is 0.64 and the average critic rating is 0.68. In total, we obtained average user ratings for 16,804 movies, and of these, 1,718 had critic ratings.

[1] http://liu.cs.uic.edu/download/data/

[2] In fact, the number of user ratings is often much larger than the number of reviews.

Distribution of Ratings (Gnitars)

Gnitar Each rating in our datasets was normalized into the range $[0, 1]$. We define the "reflected" value of the normalized rating as

$$\text{gnitar} = 1 - \text{rating}.$$

Intuitively, gnitar reflects degree of user dissatisfaction. (The reason for focusing on gnitar, rather than rating, will become clearer later.)

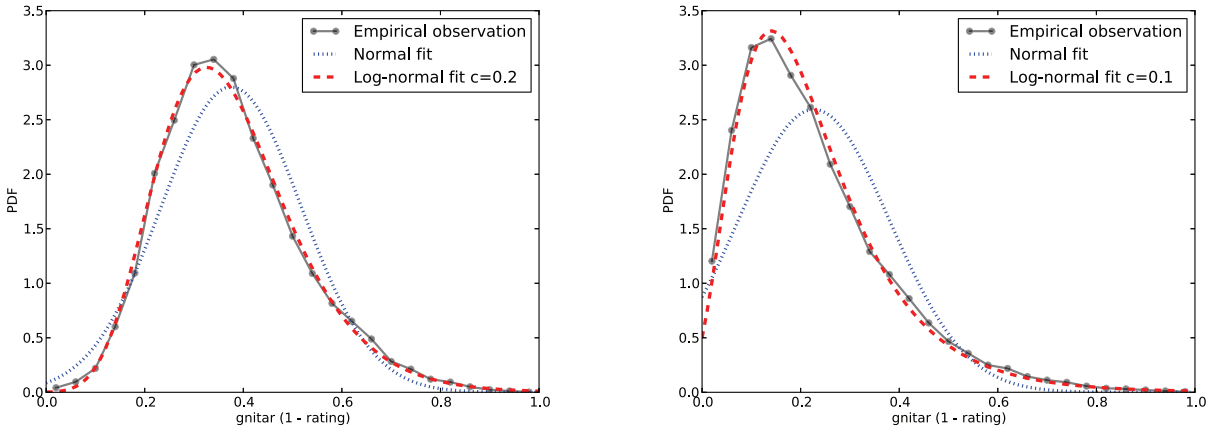
Average user gnitar of an entity

Here we focus on the average rating received by entities with at least n ratings. Ideally, the larger n is, the more accurately the average gnitar reflects expected user dissatisfaction, but extremely large n will drastically decrease the size of our datasets. We used $n = 10$ for YELP restaurants and $n = 20$ for AMAZON products.

Figure 1(a) and Figure 1(b) show the empirical distribution of average gnitar received by entities on YELP and AMAZON. We plotted the histograms (normalized so that the area under the curve is unit) of gnitar (i.e., $1 - \text{rating}$). We observe that the mean of gnitars is clearly not the midpoint of the rating scale: in both cases, on average, user dissatisfaction is lower than 0.5. As we discussed in Related Work, this is not surprising and is consistent with observations reported in previous studies (on different domains). More interestingly, we focus on the shape of the distribution, which turns out to be not symmetric around the mean. We explore this in further detail in this section.

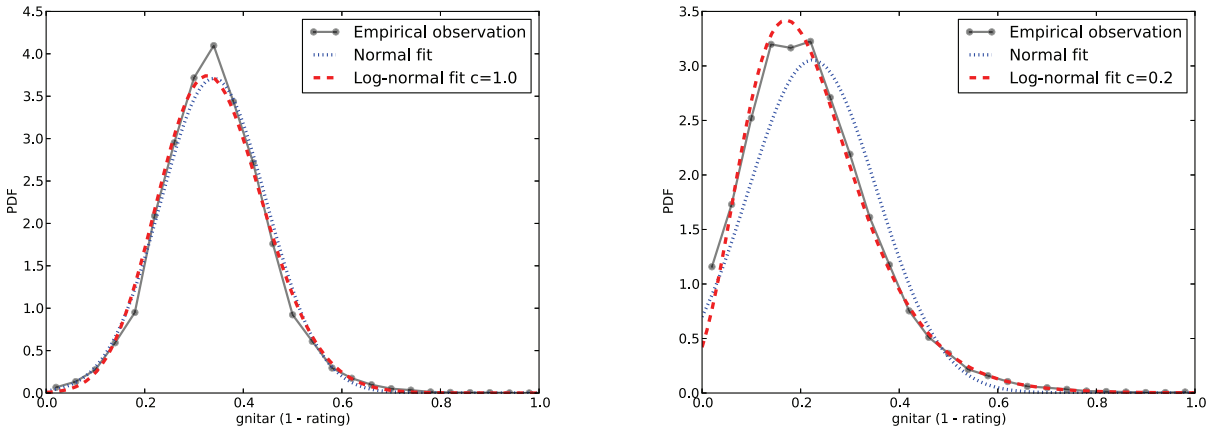
Given the central limit theorem, one would expect the average gnitars to be normally distributed. We computed the best fit to normal distribution using maximum-likelihood estimation (MLE). As we can see from Figures 1(a) and 1(b), the empirical distributions are clearly skewed away from the normal distribution fit, and the skew is more pronounced in the AMAZON dataset.

Instead, the shape of the distribution closely resembles a log-normal distribution—more specifically, a *shifted log-normal distribution*. Recall that if X is log-normally distributed, i.e., $X \sim \text{Log-}\mathcal{N}(\mu, \sigma^2)$, then $\log(X)$ has a normal distribution $\mathcal{N}(\mu, \sigma^2)$, and $Y = X - c$ has a shifted log-normal distribution with support $(-c, +\infty)$. For a given c, we can compute the best fit for shifted log-normal distribution using MLE. As can be seen from Figures 1(a) and 1(b), the empirical distributions fit shifted log-normal distribution nicely. (Note that since log-normal distributions are always "heavy" on the left-side, if we were looking into the distribution of rating, we would have to work with reflected (i.e., the mirror image of) log-normal distributions. We chose to work with gnitar to simplify the fit.)



(a) YELP (b) AMAZON

Figure 1: Distribution of average gnitar (i.e., 1 - rating) of entities.



(a) YELP (b) AMAZON

Figure 2: Distribution of average gnitar (i.e., 1 - rating) of users.

We also tried using smaller cut-off values on the number of ratings per entity, and the deviation from the normal distribution could still be observed clearly.

Average entity gnitar of a user

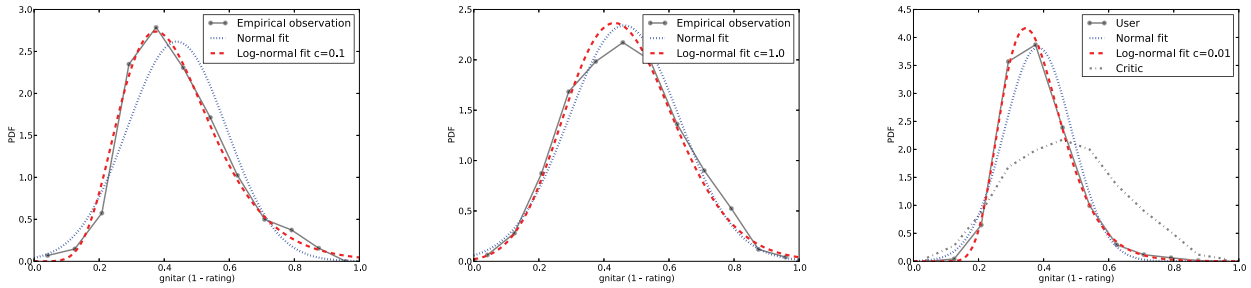
We compute the per-user average gnitar for those users who have given at least n ratings. Figure 2(a) and 2(b) show the resulting distributions.

Again, we see that the distributions are skewed away from the center in that mean $\neq 0.5$. AMAZON has a clear asymmetry, which a normal distribution does not fit, but a log-normal fits very well. YELP is more symmetric, and both normal and log-normal fit the distribution.

Average critic gnitar of an entity

We hypothesize that the skew in the average user ratings towards higher values is a result of selection bias: users select entities which they expect to like, and hence, are more probable to rate highly of the selected entities. To validate this hypothesis, we look at the distribution of average critic ratings, since critic ratings are guided by profession rather than self selection.

We examine whether average ratings from users and critics are distributed similarly or differently using the IMDB dataset. As shown in Figure 3, the distribution of average user gnitars is log-normally distributed (Figure 3(a)). In contrast, the distribution of average critic gnitars is much closer to being normally distributed (Figure 3(b)). One might suspect this difference is due to the fact that these two distributions are computed over different sets of movies. To



(a) Average user gnitar for 16,804 movies (b) Average critic gnitar for 1718 movies (c) Average user gnitar for the 1718 movies with critic ratings

Figure 3: Distribution of average gnitars of movies in the IMDB dataset: critic gnitars are normally distributed; user gnitar distributions are better captured by log-normal.

address this concern, we limit to the movies where critic ratings are available, and plot the distribution of average user gnitars for these movies, so that the histogram is supported by the same set of movies as Figure 3(b). As can be seen from Figure 3(c), we still observe a distribution more closely matched by a log-normal.

In short: critics are more normal than normal users. This provides supporting evidence that the log-normal shape is not an inherent characteristic of the quality of the entities, and it is at least partially due to characteristics of the raters. After all, critics will have less self-selection bias since it is their job to review a broad range of movies, not just the ones they believe they will like.

Another potential concern is that the difference in user and critic ratings distribution arises from critics being more *critical*, and not because of selection bias. To evaluate this, we look at users in YELP who have many ratings (at least 20). We expect these users to be experts who are more critical. Indeed, the mean of the gnitar of these users is 0.381, while the mean of average-of-all-user-gnitars is 0.370. However, when we plot the distribution of average-expert-user-gnitars, it still presents a fit to a log-normal distribution with a clear skew. These users, while more critical, are still driven by choice and not profession. This reinforces our hypothesis that the log-normal distribution is a result of selection bias, and it is not due to raters being less critical.

Temporal trends

In this section we examine the temporal aspects of the ratings using the YELP dataset.

First, we note that even if we restrict ourselves to the first 10 reviews of each entity and plot the distribution of average gnitar of these 10, the log-normal "distortion" already exists. This is quite surprising: it suggests that whatever mechanism that causes the bias happens at the very beginning of the reviewing process.

Next, we compare the average of the first 10 gnitars vs the average of the last 10 gnitars of each entity with at least 20 ratings. More specifically, for a given entity, let X_1 be the average of the first 10 gnitars, and X_n be the average of the last 10 gnitars, and let $Y = X_n - X_1$. As shown

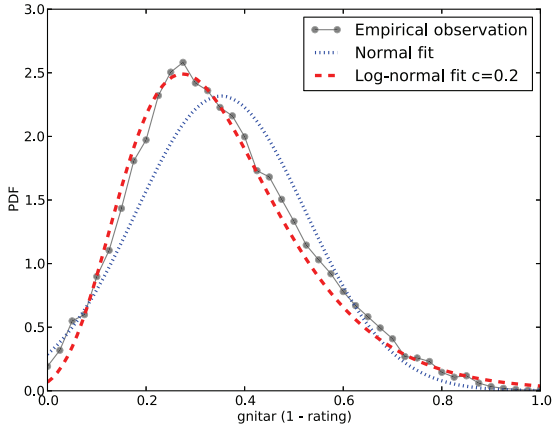


Figure 4: Distribution of average gnitar of first 10 reviews in the YELP dataset already exhibits a log-normal shape.

in Figure 5, distribution of Y is quite symmetric and has a good normal fit. Note, however, mean(Y) = 0.038541; i.e., $\text{mean}(X_n) - \text{mean}(X_1) = 0.0385$. In other words, on average, user dissatisfaction in the last 10 ratings is higher than that in the first 10.

Is this a consistent trend over time? Or should this be considered as insignificant noise? To address this question, we conducted the following experiment. Consider a moving window of size 10, where the k^{th} window consists of ratings $[k, \ldots, k+9]$. Let X_k be the average gnitar in the k^{th} window for one entity. For each k, we compute the mean of X_k over a given set of entities. Note that there is a small subtlety in this study. If we use all available entities, each k can potentially be supported by a different set of entities, where larger k will be supported by a smaller set of entities with more reviews / ratings. And since entities with more reviews tend to be more popular entities with higher ratings and lower gnitars, this could artificially cause the mean of X_k to go down as k goes up. In order to study temporal changes on the *same* set of entities, we restrict to entities

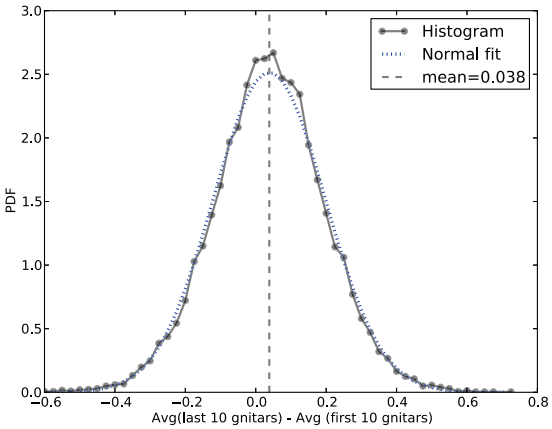


Figure 5: Distribution of the difference between the average gnitar of the last 10 reviews and that of the first 10 on the 12139 entities with at least 20 ratings in the YELP dataset.

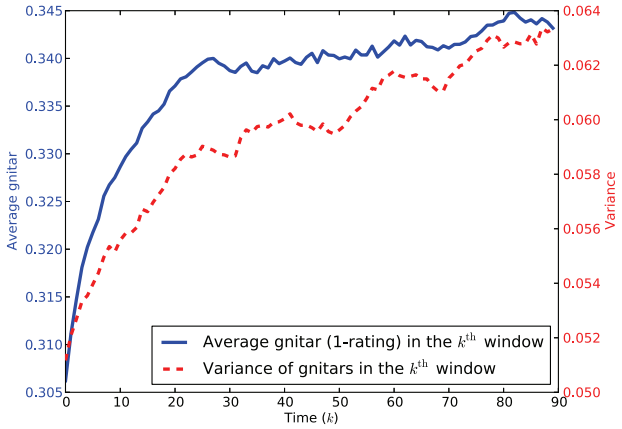


Figure 6: Mean and variance of gnitar in the k^{th} window for 5079 entities with at least 100 ratings. As k (i.e., time) increases, on average dissatisfaction (gnitar) goes up—i.e., average rating goes down, while variance of ratings goes up.

with at least 100 ratings, and plotted the mean of X_k for k values up to 90. As shown in Figure 6, on average, we observe a steady increase in gnitars (or user dissatisfaction) over time. Interestingly, if we plot the mean of V_k (where V_k is the variance of gnitars in the k^{th} window), we observe that on average, variance is also growing over time.

Note that the increasing gnitar (or decreasing ratings) is not an artifact of the reviewer base growing more critical over the years. We computed the average of all ratings from a given year in the YELP dataset, and there was no clear downward trend over the years.

Modeling Selection Bias

The main theory of selection bias is that users select entities that they expect to like, and hence, rate positively. In this section we will develop an analytical model based on this theory, which will give us an explanation of the log-normal distribution of observed average gnitars (i.e., $1 - \text{rating}$).

Let E be a set of entities and U be a set of users. We assume that each entity e in E has a *true quality*, denoted by $q(e)$. Like the transformation we did for rating to get gnitar, we will work with $q'(e) = 1 - q(e)$. In what follows, we use the term quality to refer to q' (i.e., the lower the better).[3] The true qualities of entities q' come from some probability distribution Q on the interval $[0, 1]$. Given a user u and an entity e, let $g(u, e)$ denotes the gnitar u will give to the entity e, *if* u rates e. If all users rate e, we expect the average gnitar to be proportional to its true quality, $q'(e)$.

Formally, if $\{y_i, x_i\}_{i=1}^n$ is a set of (gnitar, quality) pairs, where y_i is the user gnitar of an entity and $x_i = q'(e)$ reflects the true quality of the entity, we expect the following regression model:

$$y_i = \beta x_i + \varepsilon_i, \tag{1}$$

[3]For simplicity and ease of pronunciation, we will refrain from using ytilauq, or ʎʇilɐuɹǫ, since the meaning of quality should be clear from the context.

where ε_i is an error term. The customary assumption is that x_i and ε_i are independent, and that on average the model is correct, i.e., the mean of ε_i is 0. Thus, if we take the average of all the gnitars of a given entity, we expect it to be proportional to its true quality. Hence, the distribution of average gnitars F should following the same distribution as Q. It is not unreasonable to assume a normal distribution for Q, as it is frequently encountered in practice (For example, recall Figure 3(b). The average critic gnitar value is a reasonable proxy for the true movie quality, and it is normally distributed.)

Now we introduce selection bias into our framework, and show how it affects F. We use a simple version of the Heckmann correction model (Heckman 1979), which is a classic model to incorporate selection bias. It assumes that some observations on the dependent variable y_i are *censored*. Intuitively, each user has an a priori estimate of y_i, which denotes how likely the user will like the entity. The user only select entities where the estimate is high. In other words, there is some selection function S, such that the unit $\{y_i, x_i\}$ is censored with probability S.

Let us analyze how the selection function S changes the function F of the distribution of the observed average gnitars of entities. Let us assume that the error ε_i follows a normal distribution $\mathcal{N}(0, \sigma^2)$ with mean 0. Consider all the gnitars of an entity e with true quality given by x. Then, according to Eq. (1), the gnitars follows the distribution $\mathcal{N}(x, \sigma^2)$. The selection model says that a gnitar y is drawn from the distribution $\mathcal{N}(x, \sigma^2)$ and then discarded with probability $S(y)$. Let $t(x)$ denote the mean of the resulting probability distribution. Thus, $t(x)$ is the average gnitar an entity gets when its true quality is x.

Lemma 1. *As defined above,* $t(x)$ *is given by the expression*

$$\left(\int_0^1 y \cdot \mathcal{N}(x, \sigma^2)(y) S(y)\, \mathrm{d}y\right) / \left(\int_0^1 \mathcal{N}(x, \sigma^2)(y) S(y)\, \mathrm{d}y\right).$$

This follows from the formula for the conditional expectation of a random variable.

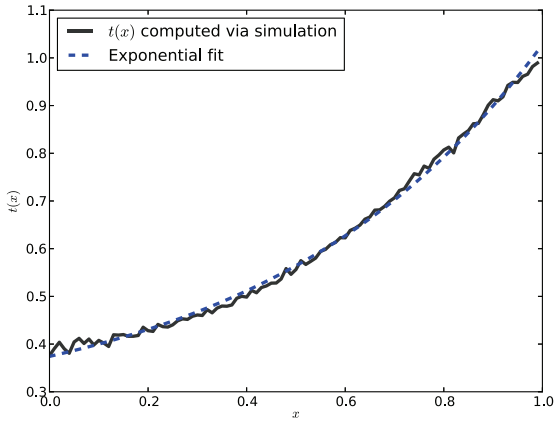


Figure 7: $t(x)$ for threshold selection function.

Tobin (1958) considers a simple form for S in his work on modeling selection bias. He considers a threshold c such that $S(x)$ is 1 for $x \leq c$ ($q \geq 1 - c$), and 0 otherwise. Let us simulate $t(x)$ using this model. We fix a threshold for the selection function, and compute $t(x)$ for different values of x by numerically evaluating the integral in Lemma 1. Figure 7 plots $t(x)$, as well as MLE fit to an exponential function $a \cdot e^{(bx+c)}$. We see that $t(x)$ has a good fit to the exponential function in this range.

Now we look at the main question, the distribution of the observed average gnitars of entities, which we denote by F. In terms of the functions defined above, the average gnitars follow the following generative model: choose x from the probability distribution Q, then transform it using function $t(x)$.

The change of variable rule says that the resulting distribution of $t(x)$ is given by

$$F = \left| \frac{\mathrm{d}}{\mathrm{d}x}(t^{-1}(x)) \right| Q(t^{-1}(x)). \quad (2)$$

Eq. (2) gives us a plausible explanation for the log-normal distribution we observed. Suppose Q follows a normal distribution, as is frequently observed in practice. If we assume an exponential form from $t(x)$, as our analysis above shows, then $t^{-1}(x)$ takes the log form, and Eq. (2) yields a log-normal distribution.

Validation

In addition to the observed log-normal distribution, we try to provide a more direct validation of our model. We look at the movie critic ratings from the IMDB dataset. We argued that since critics write reviews for movies they need to review, not just those they choose to pay to go, they do not (or to a much lesser degree) exhibit a selection bias. We consider the average critic gnitar to be the true quality of the movies. As Figure 3(b) shows, average critic gnitars are normally distributed, which is what we would expect of Q, the distribution of true quality. Next, for each x, we look at all the movies with average critic gnitar x, and plot the average user gnitar. The function is plotted in Figure 8. If we assume that critic reviews have no selection bias, and average critic gnitar is proportional to the true quality x, the curve should follow $t(x)$. We see that it indeed resembles an exponentially increasing curve, similar in shape to the curve we obtain via simulation of our model in Figure 7.

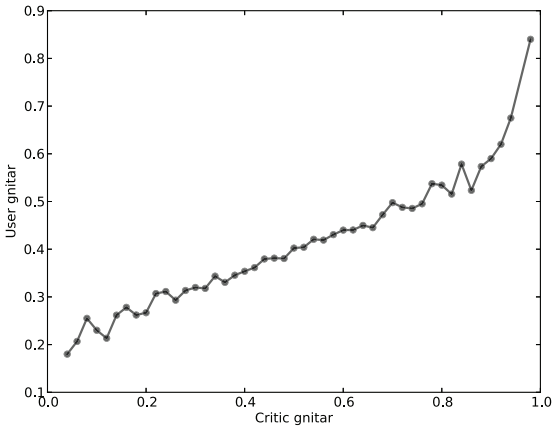


Figure 8: User gnitar vs Critic gnitar.

Other biases

Note that in addition to selection bias, there are other factors that might cause the skew towards low gnitars (i.e., favorable ratings). One example is the *choice-supportive bias*, which is the tendency to retroactively ascribe positive attributes to an option one has selected. On top of selection bias, where users select and rate entities they are likely to like, this bias can further inflate the submitted ratings. Note that our model based on selection bias allows users to be dissatisfied with entities they expect to like and rate them negatively after the choice is made (via the ε_i term); while choice-supportive bias predicts denial of unsuccessful choices. For instance, many consumers, who put lots of research and deliberation into an expensive purchase, will often refuse to admit that their decision was made in poor judgment (Cohen and Goldberg 1970). Note that the observation of mean ratings going down (average gnitars going up in Figure 6) is consistent with the selection bias model—indeed, previous work has argued that inflation in prior ratings can lead to inflated expectation in later users, and cause increasing dissatisfaction in people who chose the product as a result of the inflated expectation. Choice-supportive bias does not provide an explanation for this temporal trend.

Another factor that may increase ratings is the review spam (Jindal and Liu 2008; Li et al. 2011; Feng et al. 2012; Ott, Cardie, and Hancock 2012), where spam reviews with

high ratings can be entered for self promotion. On the other hand, spammy low ratings can also be entered to defame competitors. Systematically isolating and studying the effect of these biases is beyond the scope of this work.

The Early Prediction Problem

As an application of our model, we consider the task of predicting the eventual average rating of an entity given an initial set of ratings. In a recent paper, Yin et al. (2012) consider a related problem. They hypothesize a model where users have two personalities, *conforming* and *maverick*. The former personality prompts a user to cast her vote conforming to the majority while the later personality makes her vote different from the community. They propose a model where they assign a probability distribution over the two personalities to each user. Then, for each entity with initial ratings, they compute the predicted score each user will give to the entity, and take the average as the predicted eventual rating. However, the model has the following weakness: it assumes that every user will rate an entity in the long run. Thus, it fails to model the selection bias, which can cause the skew in the distribution of average ratings.

Here, we propose a simple model for the early prediction problem based on our model. Let $\hat{v} = (v_1, \ldots, v_k)$ be a sequence of initial $\mathrm{gnitars}$ for a given entity e. We want to predict the most likely *eventual average* gnitar of the entity, $\mathrm{gnitar}(e)$, given the observations $\hat{v}$. Let $\mathrm{gnitar}(e)$ come from a probability distribution F. Also, given an entity e with $\mathrm{gnitar}(e) = \lambda$, let the user $\mathrm{gnitars}$ for e come from a probability distribution G_λ. Thus, given F and G_λ, we want to compute the most likely $\mathrm{gnitar}(e)$ after observing the early $\mathrm{gnitars}$ $\hat{v}$.

Before we plug in the specific forms of F and G_λ that we observed, we start with a simple result when both these distributions are normal.

Theorem 1. *Suppose $F(x) = \mathcal{N}(\mu, \sigma_1^2)(x)$ and $G_\lambda(x) = \mathcal{N}(\lambda, \sigma_2^2)(x)$. Then, for any $\hat{v}$ of size k, the most likely* $\mathrm{gnitar}(e)$ *given $\hat{v}$ is*

$$\frac{\sigma_1^2 \cdot k \cdot \mathrm{mean}(\hat{v}) + \sigma_2^2 \cdot \mu}{\sigma_1^2 \cdot k + \sigma_2^2}. \tag{3}$$

Proof. The most likely gnitar is given by

$$\begin{aligned}
& \arg\max_\lambda \mathbf{P}(\mathrm{gnitar}(e) = \lambda \mid \hat{v}) \\
= \; & \arg\max_\lambda \mathbf{P}(\mathrm{gnitar}(e) = \lambda) \cdot \mathbf{P}(\hat{v} | \mathrm{gnitar}(e) = \lambda)) \\
= \; & \arg\max_\lambda \mathcal{N}(\mu, \sigma_1^2)(\lambda) \prod_{i=1}^{k} \mathcal{N}(\lambda, \sigma_2^2)(v_i) \\
= \; & \arg\max_\lambda e^{-\frac{(\lambda-\mu)^2}{2\sigma_1^2}} \prod_{i=1}^{k} e^{-\frac{(v_i-\lambda)^2}{2\sigma_2^2}} \\
= \; & \arg\min_\lambda \frac{(\lambda-\mu)^2}{2\sigma_1^2} + \sum_{i=1}^{k} \frac{(v_i-\lambda)^2}{2\sigma_2^2}.
\end{aligned}$$

This is the weighted L_2 norm, which is minimized by the weighted average given by Eq. (3). □

Eq. (3) has a nice interpretation. We take a weighted mean of the average observed gnitar and the global average. When the number of $\mathrm{gnitars}$ k is large, we prefer the entity average over the global average. Also, if the global $\mathrm{gnitars}$ have high variance (σ_1^2), we prefer the entity average. If the user $\mathrm{gnitars}$ have high variance (σ_2^2), we prefer the global average.

Now we look at our model. We know that F follows a log-normal distribution Log-$\mathcal{N}(\mu, \sigma_1^2)$, while G_λ follows a (truncated) normal distribution $\mathcal{N}(\lambda, \sigma_2^2)$.

Theorem 2. *With F and G_λ as defined above, the most likely* $\mathrm{gnitar}(e)$ *given $\hat{v}$ is*

$$\arg\min_\lambda \log\lambda + \frac{(\log\lambda - \mu)^2}{2\sigma_1^2} + \sum_{i=1}^{k} \frac{(\lambda - v_i)^2}{2\sigma_2^2}. \tag{4}$$

Since there is no closed form for the optimization, we use a simple gradient descent algorithm to compute the λ that minimized Eq. (4).

Experimental results

Based on these insights, we conduct a simple experiment on all entities with at least 50 ratings in the the YELP dataset for the early prediction problem. Recall the problem: given a sequence of initial $\mathrm{gnitars}$ $v_1, \ldots, v_k$, predict the final average gnitar of the entity. We consider the following methods: two baseline methods, a generalization of the model proposed by Yin et al. (2012), and two predictors based on our model:

- Average: use the average of $v_1, \ldots, v_k$
- Median: use the median of $v_1, \ldots, v_k$
- NCM: We generalize the naive conformer–maverick model (NCM) of Yin et al. (2012) to the non-binary case in the following manner: we compute the weight of a user u to be the cosine similarity of the user's ratings of entities to the average ratings of the corresponding entities and use these weights to compute a weighted average of $v_1, \ldots, v_k$.
- Normal based: given by Theorem 1
- Log-normal based: given by Theorem 2.

Figure 9 shows the average error values (i.e., absolute difference between the predicted average and the true average gnitar) for various values of k. Clearly, the errors decrease with increasing k for each of the predictors; this is to be expected since the predictors have more information. Median has the worst performance since its value is limited to the integers used in the raw ratings. The performances of NCM and Average are very similar (and hardly separable on the plot). For all values of k, the log-normal based predictor outperforms the normal based predictor, which outperforms all other predictors, including the NCM. As one would expect, the gap between the various predictors decreases as a function of k, with log-normal based estimator significantly outperforming the rest even for values of k as high as 10.

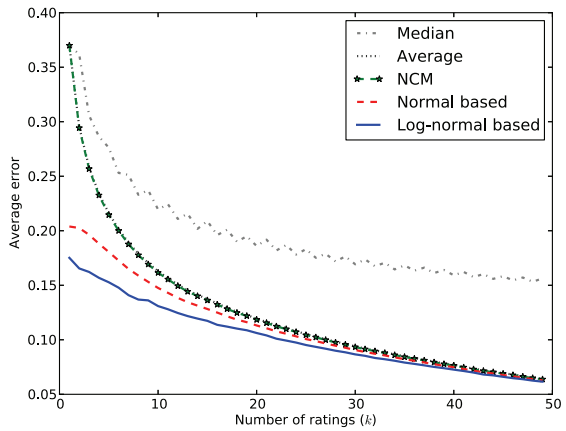


Figure 9: Error on the prediction task.

Conclusions

In this paper we study the distribution of *average* online rating of entities in a diverse range of domains. We find that in each of the three domains, the distribution clearly deviates away from the MLE fit to a normal distribution, and has a surprisingly close fit to a reflected log-normal distribution. We propose user selection bias as the underlying behavioral phenomenon, and present a simple mathematical model that gives a log-normal distribution over observed (reflected) ratings when the underlying quality distribution is normal. Experiments contrasting average critic ratings with average user ratings validated our model.

Finally, we show that a simple method derived using our selection bias model can be effective in predicting final average rating of an entity given its few initial ratings, surprizingly outperforming state of the art. Future work includes enriching the model to account for the decreasing mean of user ratings, as well as the increase in variance.

Acknowledgments

We thank Paul Resnick and Matthew Salganik for useful references.

References

Anderson, M., and Magruder, J. 2012. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal* 122:957–989.

Berinsky, A. 1999. The two faces of public opinion. *American Journal of Political Science* 43:1209–1230.

Chevalier, J., and Mayzlin, D. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* 43:345–354.

Christopher, A.; Resnick, P.; and Zeckhauser, R. 1999. The market for evaluations. *American Economic Review* 564–584.

Cohen, J. B., and Goldberg, M. E. 1970. The dissonance model in post-decision product evaluation. *Journal of Marketing Research* 7(3):315–321.

Dellarocas, C.; Gao, G.; and Narayan, R. 2010. Are consumers more likely to contribute online reviews for hit or niche products? *Journal of Management Information Systems* 27(2):127–158.

Dubin, J., and Rivers, D. 1989. Selection bias in linear regression, logit and probit models. *Sociological Methods & Research* 18(2-3):360–390.

Eliashberg, J., and Shugan, S. M. 1997. Film critics: Influencers or predictors? *The Journal of Marketing* 61:68–78.

Feng, S.; Xing, L.; Gogar, A.; and Choi, Y. 2012. Distributional footprints of deceptive product reviews. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*.

Godes, D., and Silva, J. C. 2012. Sequential and temporal dynamics of online opinion. *Marketing Science* 31(3):448–473.

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47(1):153–161.

Hu, N.; Liu, L.; and Zhang, J. J. 2008. Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. *Information Technology and Management* 9(3):201–214.

Hu, N.; Pavlou, P. A.; and Zhang, J. 2006. 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*, 324–330.

Hu, N.; Zhang, J.; and Pavlou, P. A. 2009. Overcoming the J-shaped distribution of product reviews. *Communications of the ACM* 52(10):144–147.

Jindal, N., and Liu, B. 2008. Opinion spam and analysis. In *Proceedings of the 1st ACM Conference on Web Search and Data Mining*, 219–230.

Kadet, A. 2007. Rah-rah ratings. *SmartMoney Magazine* 116.

Kramer, M. 2007. Self-selection bias in reputation systems. *Trust Management* 255–268.

Lerman, K., and Hogg, T. 2010. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th International World Wide Web Conference*, 621–630.

Li, X., and Hitt, L. 2008. Self-selection and information role of online product reviews. *Information Systems Research* 19(4):456–474.

Li, F.; Huang, M.; Yang, Y.; and Zhu, X. 2011. Learning to identify review spam. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 2488–2493.

Luca, M. 2011. Reviews, reputation, and revenue: The case of Yelp.com. Technical Report 12-016, Harvard Business School NOM Unit Working Paper.

Ma, X., and Kim, S. 2011. Revisiting self-selection biases in e-word-of-mouth: An integrated model and Bayesian estimation of multivariate review behaviors. In *Proceedings of the International Conference on Information Systems*.

Marlin, B. M., and Zemel, R. S. 2009. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the 3rd ACM Conference on Recommender systems*, 5–12.

Marlin, B. M.; Zemel, R. S.; Roweis, S. T.; and Slaney, M. 2007. Collaborative filtering and the missing at random assumption. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 267–275.

Moon, S.; Bergey, P. K.; and Iacobucci, D. 2010. Dynamic effects of movie ratings on movie revenues and viewer satisfaction. *Journal of Marketing* 74:108–121.

Ott, M.; Cardie, C.; and Hancock, J. 2012. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st International World Wide Web Conference*, 201–210.

Pang, B., and Lee, L. 2008. *Opinion Mining and Sentiment Analysis*. Now Publishers.

Pinto, H.; Almeida, J. M.; and Gonçalves, M. A. 2013. Using early view patterns to predict the popularity of YouTube videos. In *Proceedings of the 6th ACM Conference on Web Search and Data Mining*, 365–374.

Reinstein, D. A., and Snyderz, C. M. 2005. The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. *Journal of Industrial Economics* 53:27–51.

Sikora, R., and Chauhan, K. 2011. Estimating sequential bias in online reviews: A Kalman filtering approach. *Knowledge-Based Systems* 27:314–321.

Szabo, G., and Huberman, B. 2010. Predicting the popularity of online content. *Communications of the ACM* 53:80–88.

Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26(1):24–36.

Wanderer, J. J. 1970. In defense of popular taste: Film ratings among professionals and lay audiences. *American Journal of Sociology* 76(2):262–272.

Wu, F., and Huberman, B. 2010. Opinion formation under costly expression. *ACM Transactions on Intelligent Systems and Technology* 1(1):5.

Yin, P.; Luo, P.; Wang, M.; and Lee, W.-C. 2012. A straw shows which way the wind blows: Ranking potentially popular items from early votes. In *Proceedings of the 5th ACM Conference on Web Search and Data Mining*, 623–632.