The Power of Silence: An Analysis of the Aggregation and Reporting Biases in User-Generated Contents

Hongyu Chen	California State University at long Beach
Eric Zheng	University of Texas at Dallas
Yasin Ceran	University of Texas at Dallas

Abstract

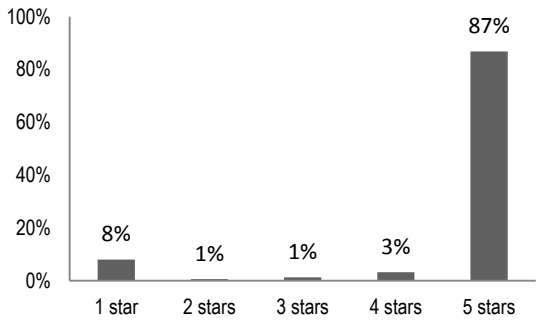
User-generated contents (UGC) such as online reviews are inherently incomplete since we do not capture the opinions of users who do not write a review. These silent users may have systematically different online experiences than those who speak up. These differences can be driven by users' differing sentiments towards their online experiences as well as their disposition to generate UGC. Indiscriminately aggregating UGC across different sentiment levels can lead to an aggregation bias and overlooking the silent users' opinions can result in a reporting bias. We develop a method to rectify these two biases simultaneously through an inverse probability weighting (IPW) approach. In the context of users' movie review activities at Blockbuster.com, we found that the average probability for a customer to post a review is 0.06 when the customer is unsatisfied with the movie, 0.23 when indifferent, and 0.32 when satisfied. A user's reporting probability with positive experience first order stochastically dominates the one with negative experience. We then adjust common UGC measures such as review volume and sentiment using these estimated reporting probabilities as weights. We show that these rectified measures yield superior predictive power, as opposed to the raw ones. Our proposed approach provides a realistic solution for business managers to paint a more complete picture from the inherently incomplete UGC.

Keywords: reporting bias, aggregation bias, user-generated contents, inverse probability weighting

1. INTRODUCTION

On September 22, 2009, Shiva Rajaraman, the product manager of YouTube, puzzled by the overwhelmingly positive ratings on most YouTube videos, made the following comment: "*[It] seems like ... Great videos prompt action [to rate videos]; anything less prompts indifference [not to rate]… if the majority of videos are getting five stars, how useful is this system really?*" Figure 1, as posted in his blog, depicts the histogram of videos at different levels of star ratings that these videos received.[1]

Figure 1: Percentage of YouTube videos (y-axis) against their star ratings (x-axis)



This observation, that almost 9 out of 10 ratings were "excellent" at YouTube, triggered a chain of reactions. Jeremy Stoppelman, the CEO of the online review site, Yelp.com, pointed out that his reviewers were less extreme though. The average rating there was 3.8 out of 5, based on a total of seven million reviews at Yelp as of 2009[2]. Although 85% of the entries were either neutral or positive, and fewer entries were negative (15%), Jeremy believed they were still trustworthy and representative. He argued that bad experience was scarce in real life, "*More often than not we are patronizing three and four-star businesses. Think back: how many one-star experiences have you had this month?*"

In reaction to this debate, Yelp.com claimed that they did not lose faith in their review system and opted to tinker their system to "*factor in the number of reviews, whether they come from experienced Yelpers or first-time reviewers, and whether those reviews were voted helpful*[3]." YouTube.com, on the other hand, simplified their 5-star rating system with a dichotomous "*Like / Don't Like*" model. Both Yelp and YouTube's cases point to the inadequacies of the current review systems.

The fundamental question behind this debate is "do online reviews faithfully represent users' opinions, and if not, what triggers someone to leave a review?" As users voluntarily contribute to online reviews, or more broadly user generated content (UGC), they have the freedom to either express their opinions, or keep silent. This self-selected behavior introduces two distinct biases into UGC samples: *an aggregation bias* and *a reporting bias*.

[1]See "http://youtube-global.blogspot.com/2009/09/five-stars-dominate-ratings.html", all URLs are verified on May 23, 2012.
[2]See "http://officialblog.yelp.com/2009/09/yelp-vs-youtube-a-ratings-showdown.html".
[3]See "http://venturebeat.com/2009/10/12/how-yelp-deals-with-everybody-getting-four-stars-on-average/".

The aggregation bias arises when one gauges UGC through aggregate measures, such as the commonly used *volume* and *sentiment*[4]. This bias occurs because the underlying UGC distribution (e.g., its specific functional form) may vary at different levels of users' opinions (sentiments). Constructing an aggregate measure without considering these different underlying distributions can result in biased measures. Take online reviews at Amazon as an example. Hu et al. (2009) find that the distribution of a book's reviews–the histogram of the count of reviews across different sentiment levels –usually exhibits a J-shape[5], with positive reviews having a disproportionally high propensity to be left by users. Similarly, Jabr and Zheng (2011) demonstrate that the specific functional form of this distribution changes drastically across books at different sentiment levels. To one extreme, if a book receives two dichotomous types of reviews, with half being extremely positive and half being extremely negative, the overall aggregated sentiment becomes neutral, which clearly does not reflect the polarity of the reviews, and therefore is biased as a result of aggregation.

A concomitant bias is the reporting bias, where users with different opinions tend to have different propensities to report (i.e., to produce UGC). For example, eBay documented that 99% of its user feedback was positive (Dellarocas and Wood 2008). However, this does not imply eBay's exuberant success; rather it is more likely a result of users' unwillingness to express their negative shopping experiences. As a result, the reported UGC is intrinsically biased[6].

Although in this paper we examine the aggregation bias and the reporting bias in the context of online reviews, these two biases are prevalent in other forms of UGC, such as online forums, blogs, tweeter etc. We specifically address two research questions:

1. *Are there any systematic biases in UGC? If there are, how can we accurately quantify them?*
2. *How can we rectify aggregation and reporting biases and what are the economic implications of doing so?*

The key intuition behind these two questions is that a user's tendency to report depends critically on her sentiment toward her online (shopping) experience. In analyzing UGC, we thus need to first fully understand the user's reporting probability at a particular sentiment level. This underpins our approach to tackle with these two biases.

[4]In the context of online reviews, volume is measured as the total number of reviews a product receives; and sentiment (or valence, ratings) is measured as the average ratings of these reviews (Duan et al. 2008a).

[5] The J-shape indicates that there exist a large number of five-star reviews for a product, followed by some one-star and very few mediocre reviews.

[6]Our defined reporting bias is closely related to the underreporting bias in statistics, but with subtle differences. The classical underreporting problem (Winkelmann1996, Fader and Hardie 2000) arises when an individual fails to report all of the events, e.g., a worker tends to report fewer (i.e., underreport) absent workdays than actual. In this sense, UGC also represents an underreported sample of user experiences (without those of the silent users). We choose to use a different term, reporting bias, to allow a user's underreporting behavior to be a function of her sentiment (i.e., to a priori allow the case when a user may or may not underreport at certain sentiment level)..

The challenge is that, however, we cannot observe a user's underlying decision making process regarding whether to express her opinion through UGC. As a result, once the "silent" user decides not to report, we cannot capture any information (such as sentiment) on her. Taking online review of movies as an example[7], in order to model a user's reporting behavior, it is necessary to know how much she likes (or dislikes) a movie and whether she posts a review on that. The dilemma is that, unless a user writes a review, in most scenarios, the researcher is unable to observe whether that user has watched the movie or not, let alone her specific opinion toward the movie. Thus, what is observed in online reviews is naturally incomplete. The crucial task here is how to portray a complete picture from incomplete data by accounting for the reporting bias.

Moreover, when dealing with the reporting bias, a further complication is that this bias is interwoven with the aggregation bias. At first glance, one may think that the aggregation bias can be easily tackled with by separating reviews into different piles, one for each level of sentiment of interest. However, the reviewers of each pile may not be homogeneous, and their tendency to review may differ. Thus UGC measures constructed in this manner by aggregating across heterogeneous reviewers is still biased.

In this paper, we develop an integrated stochastic model to directly model the underlying data generating process of UGC. To mitigate the reporting bias from the partially observed UGC (e.g., reviews posted), we resort to the rich statistical and marketing literature on modeling customer's repeat buying behavior (Fader and Hardie 2000, Goodhardt et al. 1984, Winkelmann 2008). Specifically, we build on the classic Beta-Binomial/Negative Binomial Distribution (BB/NBD) framework as applied in Fader and Hardie (2000) and Winkelmann (2008). Using the DVD rental business at Blockbuster as an illustration, this BB/NBD process prescribes that a random customer's rental rate (i.e., how often she is to rent movies) follows a NBD process; and after each rental, she decides whether or not to post a review following a Beta-Binomial process. To account for the fact that a user's sentiment level may affect her reporting decision, we split a user's reporting process into separate sub-processes, each for a different sentiment level, and distributed as BB/NBD.

To evaluate our model, we acquired a unique dataset from Blockbuster.com. This dataset documents each customer's review history as well as her complete rental transactions, including those rented movies that she did not review. This information enables us to evaluate our model performance and account for these silent users. Importantly, we only use this rental information to evaluate our model; we do not use this information (which is often unavailable to researchers) during the course of model building.

After validating the model, we investigate user's reporting behavior. We found that the average probability for a customer to post a review is 0.06 when the customer is unsatisfied with the movie, 0.23

[7]We obtained an online movie rental dataset from Blockbuster (see Section 3) and we will use this as a running example throughout this paper, without loss of generality.

when indifferent, and 0.32 when satisfied. Overall, the reporting probability when the user is positive first order stochastically dominates the reporting probability when the user is negative.

With the estimated reporting probabilities, we then adopt the inverse probability weighting (IPW) approach (Wooldridge 2010) to de-bias the commonly used UGC measures - including volume and sentiment - where the inverse of the inferred reporting probabilities are used as the weights. The IPW-adjusted measures are then applied to replicating an econometrics model that investigates the effect of volume and sentiment on sales, as specified in Forman et al. (2008). Using the aforementioned Blockbuster data, we compare the scenario using conventional UGC measures against the one using our IPW adjusted measures. We found that the effect of volume on sales tends to be exaggerated while that of sentiment tends to be underestimated, when the reporting bias is unaccounted for. This suggests that our conventional wisdom on UGC may be misleading.

This study makes the following key contributions to the UGC literature:

1. We quantify the magnitude of the aggregation and reporting biases in UGC, by inferring the reporting probabilities for different types of customers.
2. We propose a model that directly account for the data generating process of these two biases. The model draws on but extends the well-established BB/NBD framework.
3. We propose new approaches to validating our model using the real data from Blockbuster.
4. We develop an inverse probability weighting (IPW) approach to rectify the reporting errors. We empirically demonstrate the substantial improvement of econometrics analyses after controlling for these biases.

The rest of the paper is organized as follows. Section 2 reviews the literature. Section 3 introduces the research context and describes the dataset to be used. Section 4 develops the model to address the two biases and Section 5 presents the experimental results to evaluate the proposed model. Section 6 demonstrates the application of the model through an econometric analysis, and Section 7 concludes the paper.

2. LITERATURE REVIEW

Our proposed model attempts to simultaneously address the aggregation and reporting biases prevalent in the UGC domain. We review two relevant streams of literature: one on the economic value of UGC and the other on the treatment of these two biases in statistics and econometrics.

2.1. User Generated Contents

The majority of studies on the economic value of UGC attempt to establish a link between UGC measures (e.g., volume of UGC) and economic measures (e.g., product sales) (Chen et al. 2008, Duan et al. 2008b). The two most common (aggregate) measures of UGC are volume and sentiment, where the

former is a simple count of the number of UGC posts and the latter represents the average opinion (e.g. positive or negative) of UGC users. In the context of online reviews, volume has been predominantly found to exert a positive effect on sales (Chen et al. 2008, Chevalier and Mayzlin 2006, Duan et al. 2008a, Liu 2006). Findings on sentiment, however, are mixed, with some documenting a positive effect on sales (Chen et al. 2008, Clemons et al. 2006); others showing no significant impact (Duan et al. 2008a, Forman et al. 2008, Mudambi and Schuff 2010). Besides volume and sentiment, several other UGC measures have also been proposed, which include dispersion of sentiment (Clemons et al. 2006), helpfulness of a review (Mudambi and Schuff 2010), identity of reviewers (Forman et al. 2008), and readability of a review (Ghose and Ipeirotis 2011).

Several recent studies have argued that more granular analysis is needed in lieu of these aggregate measures. Etzion and Awad (2007) find that volume has a positive effect on sales of products only when the valence (average rating) is perceived to be positive by customers, while volume turns into having a negative effect on sales in the case of negative valence. They conclude that pooling reviews altogether may not help correctly capture the relationship between volume and sales. In the same line of reasoning, researchers have started to separate reviews based on more granular level, e.g., at different levels of user sentiments. Chevalier and Mayzlin (2006) find that 1-star reviews have a higher impact than 5-star reviews, as do Basuroy et al. (2003). Clemons et al. (2006) define the variance of ratings and show that only the most positive quartile of reviews plays a significant role on sales growth. All these studies point to the existence of an aggregation bias.

2.2. Treatments of the Aggregation Bias

In practice, aggregating granular data into higher levels is common and oftentimes necessary. However, this aggregation process itself may introduce bias. Fader and Hardie (2010) report a case that aggregation can cause errors in the estimation of customer lifetime value. Kelejian (1995) discusses why aggregation bias might occur in the logit model and how to test the existence of such biases. In some cases, aggregation at different granular levels may lead to contradictory inferences, as illustrated by the classic Simpson's paradox (Simpson 1951). Bickel et al. (1975) reported an interesting lawsuit against UC Berkeley for its alleged discrimination against women in admission, where the acceptance rate of female applicants was much lower than that of males (35% vs. 44%) at the university level. However, when drilling down to the department level, it was discovered that almost every department was favoring female students. According to Bickel, the best explanation for the contradiction is a behavioral one that women tend to apply for more competitive departments. In the context of UGC, we similarly find that the number of positive reviews tends to be inflated while that of the negative ones deflated. Aggregation errors can arise when we sum up the raw reviews without considering the different reporting probabilities of users with different opinions.

That aggregation may lead to erroneous analysis has also been studied in other contexts in IS. Abhishek et al. (2011) show how search engines are adversely affected due to the aggregation bias. Jabr and Zheng (2011) discuss the aggregation bias and propose the use of more granular data to alleviate this problem. However, since the aggregation bias is interwoven with the reporting bias in UGC, researchers need to address the aggregation bias and the reporting bias simultaneously.

2.3. Treatments of the Reporting Bias

Our defined reporting bias problem differs from the traditional self-selection problem in econometrics. Namely, it cannot be readily addressed by the Heckit approach (Heckman 1979, Wooldrige 2010), a common approach for handling the self-selection bias. The common scenario for the self-selection bias assumes that which user self-selects into the sample is observable to researchers. In the classical married women's labor supply example (Wooldridge 2010, p.807), researchers know which worker voluntarily decides (self-selects) on whether to join the work force or not. And this is precisely the feature that makes the first stage of the Heckit (which normally models the self-selection probability as a Logit or Probit function of other exogenous variables) plausible (Wooldridge 2010, p.806). In our case, we simply do not know who these silent users are.

Other treatments proposed in the IS literature for self-selection bias in UGC do not fully address the reporting bias either. For example, Li and Hitt (2008) discuss one special type of self-selection bias in a review system. They consider the case when a customer's preference towards an author affects the customer's timing of purchasing on that author's new book. Early adopters may exhibit a positive bias, which distorts the online reviews at an early stage and can mislead new customers. Their study focuses on reported reviewers without regard to silent customers.

The pioneering study of Dellarocas and Wood (2008) on eBay's feedback mechanisms is related to our work. They find that eBay traders are more likely to post feedback when satisfied than otherwise. They propose a logit model to estimate the reporting probability of each eBay transaction based on participating traders' historical feedbacks. However, in their transactional level logit model, it is necessary for the researcher to observe all the transactions, including those of the silent users. This cannot be generalized to other UGC settings for the same reason we discussed for the Heckit approach. Further in their population level model, reporting probabilities are treated uniformly across all traders, i.e., each trader's probability is effectively equal to the population mean. We adopt a stochastic modeling approach to allow *user-specific* reporting probabilities. This enables us to establish the stochastic dominance among the reporting probabilities under different scenarios (sentiment levels).

In another closely related work, Hu et al. (2009) find that the histogram of online reviews across different sentiment levels exhibits a J-shape pattern. However, their lab experiments show that the real pattern of sentiments should approximate a normal distribution. To rectify the biases caused by the

(misrepresenting) J-shape pattern, researchers should regress on not only the mean of the ratings but also on the variance of the ratings. We advance Hu's effort by quantifying users' reporting probabilities through an integrated stochastic model.

Our technique of treating the reporting bias roots in a family of count models on underreporting in statistics. Underreporting refers to a broad class of phenomena that the reported number of events is less than the actual total number of events. For example, a worker tends to report fewer absent workdays than actual when filling her work status report (Allen 1981, Barmby et al. 1991). Winkelmann (1996) conducts a Markov Chain Monte Carlo analysis to augment a Poisson regression to recover the underreported days. Fader and Hardie (2000) extend the model to customers' repeated purchase case, where customers may not be able to recall and report all their purchases during a period. The authors adopt the well-established BB/NBD framework to account for underreporting. In our case, a customer not only underreports, but also selectively reports her experience, namely her reporting probability depends on her sentiment. In Section 4, we build an integrated model to incorporate both underreporting and selective reporting processes.

Since those silent users' real events are not observed, directly comparing the estimated number of events (e.g., via the BB/NBD model) with the real one is often not possible. Evaluation of an underreporting model is usually done indirectly, for example, by comparing the expected number of absent days occurring in the future period given that a worker has been absent for x number of days during the period of observation (Greene 1982, Morrison and Schmittlein 1981). We will propose several novel ways to validate our model in Section 4.

3. RESEARCH CONTEXT: THE BLOCKBUSTER ONLINE DVD RENTAL BUSINESS

In order to evaluate our model to be built, preferably we need to know the (reviewing) behaviors not only for the users who choose to post a review, but also for the users who remain silent. We obtained such a unique dataset from Blockbuster's online movie rental business. The online movie rental industry is a subscription-based business serving a group of online subscribers. Because customers have to first become "members" in order to rent movies online from Blockbuster, Blockbuster's customer base is relatively stable compared to other online types of retailing businesses. Additionally, in this industry, each subscriber's rental and review history is readily tracked. At Blockbuster.com, after renting and (presumably) watching a movie, a customer forms her opinion (sentiment) towards the movie and then decides whether or not to post a review on it[8]. The company displays the reviews of the users and the average ratings of the movies as reviews are being entered.

[8] In our dataset, only 2.6% of the reviews are posted before renting.

This dataset is unique in that it contains the detailed records of both rental transactions and review activities for 201,258 active online subscribers, i.e., we know who those silent users (of a movie) are. This enables us to validate our model directly. In this dataset, we track all the customer reviews that are posted between October 2009 and February 2010 for 69,786 movies. The dataset also provides the time stamp for each movie rental and each review. In Table 1 below, we present the descriptive statistics.

Table 1: Summary statistics

Variable	N	Mean	Std. Dev.	Min	Max
Rental/User	201258	50.73	31.57	2	373
Review/User	201258	9.82	25.65	0	2211
Negative Review/User	201258	1.23	6.78	0	2087
Neutral Review/User	201258	3.00	9.01	0	562
Positive Review/User	201258	5.59	15.01	0	929
Rental/Movie	69786	146.31	1933.89	1	234227
Review/Movie	69786	28.32	353.51	0	23725
Negative Review/Movie	69786	3.56	46.39	0	3834
Neutral Review/Movie	69786	8.64	105.54	0	5752
Positive Review/Movie	69786	16.12	229.46	0	19163

At Blockbuster.com, customers rate a movie with a scale ranging from 5 (negative) to 50 (positive), with a minimal increment of 5. Blockbuster treats a review as negative when the rating is in the range of 5-20, neutral when 25-30, and positive when 35-50. As is to be expected, the rating data is skewed towards positive. As it's clearly shown in Table 1, since the number of negative reviews is less than the neutral ones, the sentiment data is not J-shaped as described by Hu et al. (2009).

4. MODEL DEVELOPMENT

In this section, we develop our integrative stochastic model that directly considers the UGC data generating process, i.e., customers' underlying rental and reviewing behaviors altogether.

The canonic story behind this model is that a random customer first makes a purchasing (rental) decision (e.g., renting a movie from Blockbuster); she then forms a specific opinion (sentiment) towards the purchased product. Without loss of generality, we consider three types of sentiments: positive, negative and neutral. Specifically, we denote the probability that the user likes the movie as q_1, that she feels neutral about it as q_0, and that she dislikes it as q_{-1}. Then the user's purchasing process is decomposed into three separate stochastic sub-processes, depending on her sentiments on the movies.

Next the customer decides whether or not to report her opinion via posting a review. It is in our interest to examine whether the user has different reporting probabilities at different levels of sentiments. We thus separately denote her specific reporting probabilities as p_{-1} when her sentiment is negative, p_0 when neutral, and p_1 when positive. Below we present the model.

4.1. Modeling Users' Purchasing Behavior

Our model begins with the following assumptions:

1. A customer's purchases follow a Poisson process with the purchase rate λ, during a unit period of time. The Poisson process is standard in modeling customer's purchasing behavior when her purchase rate is stable (Morrison and Schmittlein 1988). This stationary assumption specifically fits well to the Blockbuster's business, since their service is subscription based with quota limitations[9] and the rental rate per month is stable.
2. For each movie a customer rents, we assume that with probability q_{-1} she dislikes the movie after watching it; with q_0 she feels neutral about it, and with q_1 she likes it, where $q_{-1} + q_0 + q_1 = 1$. We assume that the set of three probabilities stays constant for the customer during the research time span. In a sense, the composition of (q_{-1}, q_0, q_1) reflects a customer's capability of making correct choices, because a priori, a customer intends to only rent good movies. It is reasonable to assume that her capability does not fluctuate much in a relative short period of time[10].

Given the above two assumptions, a customer's purchasing process can be broken down into three independent Poisson processes: F_{-1}, F_0, and F_1, with purchase rates $\lambda_{-1} = \lambda \cdot q_{-1}$, $\lambda_0 = \lambda \cdot q_0$ and $\lambda_1 = \lambda \cdot q_1$, respectively. This is known as the random split of a Poisson process, proof of which is provided in Ross (2003). For ease of exposition, throughout this paper we use subscript i to index the sentiment level i, $i = -1, 0, 1$.

Some customers rent very frequently, others hardly at all. To capture the heterogeneity of purchasing rate, we assume that λ_i $(i = -1, 0, 1)$ in the population follows a gamma distribution, a standard conjugate prior to the Poisson distribution. That is, $f(\lambda_i) = \alpha_i^{r_i} \lambda_i^{r_i - 1} e^{-\lambda_i \alpha_i} / \Gamma(r_i)$ with shape parameter r_i and scale parameter α_i. Gamma, combined with the previously assumed Poisson rental rate, yields the classic NBD model (Winkelmann 2008). Dunn et al. (1983, p.256) have shown, through empirical studies and simulated experiments, that the NBD assumptions are reasonable for modeling buyers' behaviors. In our case, the aggregate number of purchases under each of the three processes, F_{-1}, F_0, and F_1, follows a distinct NBD process, the pdf of which for F_i is as

[9]Blockbuster offers three types of plan: $9.99, $14.99 or $19.99 per month during our data collection period. A user could check out 1, 2 or 3 DVDs per time respectively. See "https://www.blockbuster.com/signup/m/plan".

[10]It may be argued that a user's experience depends on the supply of high quality movies, e.g., movies in summer are better than those in other seasons, so does a user's chance of renting good movies. Thus, the probabilities (q_{-1}, q_0, q_1) will fluctuate over time. However, unlike in the case of movie theaters where the number of good movies being shown in theatres is limited, users at Blockbuster could select from a large and relatively stable collection of movies (69,786 titles in our dataset), which effectively alleviates quality fluctuation. Further, our data spans from Oct. 2009 to Feb. 2010. Seasonality of the quality of new movie released is not a severe concern here. Lastly, it is important to simplify our model to maintain model tractability. Our model's good predictability (as shown in Figure 3) partially validates our model assumptions.

$$P(N_i = n_i|r_i, \alpha_i) = \frac{\Gamma(r_i + n_i)}{\Gamma(r_i)n_i!}\left(\frac{\alpha_i}{\alpha_i + 1}\right)^{r_i}\left(\frac{1}{\alpha_i + 1}\right)^{n_i}, \qquad i = -1, 0, 1, \tag{1}$$

where n_i is the total number of purchases of the customers for process F_i. $\Gamma(\cdot)$ is the gamma function, which is defined by $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ (Abramowitz and Stegun 1965).

4.2. Modeling Users' Reviewing Behavior

After watching a movie, a customer decides whether to post a review or not. Our assumption 2 implies that customers behave consistently: every time a customer likes a movie, she will post a review with probability p_1; when she is neutral, the probability is p_0; and when she dislikes a movie, her reporting probability is p_{-1}. Thus, this customer's propensity to write a review when being positive is simply a Bernoulli choice for the Poisson process F_1, so are the propensities for the negative and the neutral cases. In other words, she makes independent, binary decisions every time. We denote a customer's total number of positive reviews as x_1 (x_0 and x_{-1} can be defined accordingly). For sentiment level i, the conditional probability of the total number of reviews, x_i, given the total number of rentals, n_i, and reporting probability, p_i, follows a binomial distribution:

$$For\ Process\ F_i: \quad P(X_i = x_i|n_i, p_i) = \binom{n_i}{x_i} p_i^{x_i}(1 - p_i)^{n_i - x_i}, \quad i = -1, 0, 1. \tag{2a}$$

Again, some users may review every movie they watched, others never do. We thus model this heterogeneity by allowing the reporting probabilities p_i to differ across individuals, using the natural conjugate prior - the beta distribution - for the binomial distribution:

$$g(p_i) = p_i^{a_i - 1}(1 - p_i)^{b_i - 1}/B(a_i, b_i), \tag{2b}$$

where a_i and b_i are shape parameters of the Beta distribution. $B(\cdot)$ is the beta function, which is defined as $B(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}dt$ (Abramowitz and Stegun 1965). Equations (2a) and (2b) combined lead to the Beta-Binomial (BB) model (Chatfield and Goodhardt 1970)[11]:

$$P(X_i = x_i|N_i = n_i) = \binom{n_i}{x_i}\frac{B(a_i + x_i,\ b_i + n_i - x_i)}{B(a_i, b_i)}, \quad i = -1, 0, 1. \tag{2c}$$

Equation (2c) defines the distribution of the number of negative, neutral and positive reviews a random customer has posted when we know how many movies she had negative, neutral or positive sentiments with.

4.3. Combing Users' Purchasing and Reviewing Behaviors

[11]The Beta-Binomial model is a special binary choice case of the Dirichlet-multinomial model for multinomial choices (Fader and Schmittlein 1993, Goodhardt et al. 1984).

The purchasing process and the choice process (i.e., reviewing decision) are usually assumed to be independent in the repeat buying literature (e.g., Fader and Schmittlein 1993, Winkelmann 2008). For us, this is equivalent to assuming that the purchase rate λ_i and reporting probability p_i are independent[12]. Combining the distribution of the number of purchases derived in Equation (1) with the reporting process defined in Equation (2c), we obtain the BB/NBD model (Schmittlein et al. 1985). The marginal probability of the number of reviews at sentiment level i is

$$P(X_i = x_i) = \frac{\Gamma(r_i + x_i)}{\Gamma(r_i)x_i!}\left(\frac{\alpha_i}{\alpha_i + 1}\right)^{r_i}\left(\frac{1}{\alpha_i + 1}\right)^{x_i}\frac{\Gamma(a_i + x_i)}{\Gamma(a_i)}\frac{\Gamma(a_i + b_i)}{\Gamma(a_i + b_i + x_i)} \times {}_2F_1\left(r_i + x_i, b_i; a_i + b_i + x_i; \frac{1}{\alpha_i + 1}\right), \quad i = -1, 0, 1, \tag{3}$$

where ${}_2F_1(\cdot)$ is the Gaussian hypergeometric function (Johnson et al. 1992).

Equation (3) prescribes that a customer's number of reviews at each sentiment level (X_i) follows a BB/NBD distribution. In Equation (3), the number of a user's purchases, n_i, has been integrated out and only her number of reviews at each sentiment level (x_i) is involved. In practice, x_i is not difficult to acquire since a user's historical review data is usually available at most UGC settings (e.g., Yahoo! review, Amazon review, and YouTube rating).

4.4. Computing Users' Reporting Probability

The reporting probability of a random user, conditional on her number of past reviews could be derived as:

$$g(p_i|x_i) = \frac{{p_i}^{a_i + x_i + 1}(1 - p_i)^{b_i - 1}}{B(a_i + x_i, b_i)}\frac{(1 - (1 - p_i)/(\alpha_i + 1))^{-(r_i + x_i)}}{{}_2F_1(r_i + x_i, b_i; a_i + b_i + x_i; 1/(\alpha_i + 1))}, \qquad i = -1,0,1. \tag{4}$$

This is one of the key interests of this paper - quantifying the review propensity. By plugging in the parameters estimated from Equation (3) into Equation (4), we obtain a customer's reporting probabilities at different sentiment levels, given only her historical review profile (x_{-1}, x_0, x_1). These estimated reporting probabilities are central to de-bias the reported UGC as we will show later. Because users' historical posts are available at most UGC hosting platforms, Equation (4) is very generic.

A less generic but also very practical case is when the total number of purchases, n, is also known. We derive the distribution of the reporting probability of a user given her review history and purchase history, $g(p_i|x_{-1}, x_{0,}x_1, n)$ in equation (5), where the detailed derivation is provided in the Appendix.

$$g(p_i|x_{-1}, x_0, x_1, n) = \frac{\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\left\{\frac{\mathrm{B}(a_i+x_i+1, l+b_i)\Gamma(l+x_i+r_i)}{l!(\alpha_i+1)^l}P(N_j+N_k=n-x_i-l|X_j=x_j, X_k=x_k)\right\}}{\mathrm{B}(a_i+x_i, b_i)\Gamma(r_i+x_i)\ {}_2F_1\left(r_i+x_i, b_i; a_i+b_i+x_i; \frac{1}{\alpha_i+1}\right)P(N=n|X_{-1}=x_{-1}, X_0=x_0, X_1=x_1)} \tag{5}$$

[12] We later in Section 4.4 relax this independence assumption by allowing the reporting probability to be dependent on the total number of purchases, which is new to the BB/NBD literature.

We would like to highlight that this result delineates a scenario that is new to the BB/NBD literature. Effectively, we contribute to the literature by relaxing the assumption that the purchase process is independent of the choice process through *n*, which is assumed to be known here. This scenario does not naturally arise in the conventional BB/NBD setting. To the contrary, it is characteristic of BB/NBD to be agnostic of *n* (by integrating it out). Considering that the data input to BB/NBD is only the histogram of user reviews, its ability to 'recover' the purchase component can be questionable (Zheng et al. 2012, Fader and Hardie 2000). Knowing *n* will give the dataset more "texture" and such data augmentation beyond the simple histogram can allow the reliable estimation parameters with the ability to make direct linkages to (and inferences about) the reporting probability.

4.5. Evaluating the Model's In-Sample Performance

Though in Equation (3), n_i, the number of purchases for sentiment level i, has been integrated out, it is straightforward to infer it, conditional on the number of observed reviews. For example, given a customer's reported number of negative reviews, one can compute the number of movies lowly-rated by that customer. This is expressed in the following equation:

$$\begin{aligned} &P(N_i = n_i | X_i = x_i) \\ &= \frac{\Gamma(r_i + n_i)}{\Gamma(r_i + x_i)(n_i - x_i)!}\left(\frac{1}{\alpha_i + 1}\right)^{n_i - x_i} \frac{\Gamma(a_i + b_i + x_i)}{\Gamma(a_i + b_i + n_i)} \frac{\Gamma(b_i + n_i - x_i)}{\Gamma(b_i)} \\ &\qquad / \ {}_2F_1\left(r_i + x_i, b_i; a_i + b_i + x_i; \frac{1}{\alpha_i + 1}\right), \quad i = -1, 0, 1. \end{aligned} \tag{6}$$

Equation (6) enables us to estimate n_i, the number of movies a user likes/dislikes/feels indifferent, which is unobservable. Had we observed n_i, we would be able to evaluate our model by comparing this estimate with the actual number of rentals at each sentiment level directly. However, a silent customer's real sentiment is un-observable. We thus have to propose a feasible evaluation scheme as follows.

Following Hausman and McFadden (1984), we utilize the fact that the total number of purchases of a customer ($n = n_{-1} + n_0 + n_1$) is often observed. For example, Blockbuster knows the total number of rentals of a customer (over a certain period of time), but its breakdowns at each sentiment level are unknown. Moreover, for a customer with a review profile (x_{-1}, x_0, x_1), her total number of purchases, n, could be derived by:

$$\begin{aligned} &P(N = n_{-1} + n_0 + n_1 | X_{-1} = x_{-1}, X_0 = x_0, X_1 = x_1) \\ &= P(N_{-1} = n_{-1} | X_{-1} = x_{-1}) \oplus P(N_0 = n_0 | X_0 = x_0) \oplus P(N_1 = n_1 | X_1 = x_1), \end{aligned} \tag{7}$$

where $\oplus$ denotes the convolution operation. $P(N_i = n_i | X_i = x_i)$ on the right hand side of Equation (7) can be computed via plugging the estimated parameters of Equation (3) into Equation (6). Note that, Equation (7) is an outcome of the random split of Poisson process where the three processes (F_{-1}, F_0, F_1)

are independent of each other. We then can use Equation (7) as a benchmark to evaluate the model by comparing the observed n with the estimated one.

There is one caveat though. Equation (7) only specifies a probability distribution conditioned on a customer's specific review profile (x_{-1}, x_0, x_1). In reality, the number of negative reviews a customer has posted (x_{-1}) could vary from 0 to hundreds, so does the number of neutral reviews (x_0) and positive reviews (x_1). The number of possible combinations of (X_{-1}, X_0, X_1) could be astronomical. Comparatively, the number of customers with a specific profile (x_{-1}, x_0, x_1) is very limited, which renders the evaluation approach through Equation (7) impractical. An alternative approach to validate our model is to pool customers with different profiles together. For example, we can derive $P(N = n_{-1} + n_0 + n_1 | X_{-1} \neq 0, X_0 \neq 0, X_1 \neq 0)$, the probably of a customer's total purchase given that she has posted at least one review for each sentiment level, as follows:

$$\begin{aligned} &P(N = n_{-1} + n_0 + n_1 | X_{-1} \neq 0, X_0 \neq 0, X_1 \neq 0) \\ &\qquad = P(N_{-1} = n_{-1} | X_{-1} \neq 0) \oplus P(N_0 = n_0 | X_0 \neq 0) \oplus P(N_1 = n_1 | X_1 \neq 0), \end{aligned} \tag{8}$$

where

$$P(N_i = n_i | X_i \neq 0) = \frac{1 - \frac{\Gamma(r_i+n_i)}{\Gamma(r_i) n_i!}\left(\frac{1}{\alpha_i+1}\right)^{n_i} \frac{\Gamma(a_i+b_i)}{\Gamma(a_i+b_i+n_i)} \frac{\Gamma(b_i+n_i)}{\Gamma(b_i)} \; {}_2F_1\left(r_i, b_i; a_i + b_i; \frac{1}{\alpha_i+1}\right)}{1 - \left(\frac{\alpha_i}{\alpha_i+1}\right)^{r_i} \left(\frac{1}{\alpha_i+1}\right) \times \; {}_2F_1\left(r_i, b_i; a_i + b_i; \frac{1}{\alpha_i+1}\right)}$$
$$\times \frac{\Gamma(r_i + n_i)}{\Gamma(r_i) n_i!} \left(\frac{\alpha_i}{\alpha_i + 1}\right)^{r_i} \left(\frac{1}{\alpha_i + 1}\right)^{n_i}, \qquad i = -1,0,1.$$

Plugging in the parameters estimated from Equation (3) into Equation (8) and then comparing this empirical distribution against the real distribution from the actual rental data at Blockbuster, we have a way to evaluate the fitness of our model.

4.6. Evaluating the Model's Predictive Performance

Predicting a customer's future period purchase given her observed first period activity is often of central interest in any businesses. We further evaluate our model's predictive performance by computing the conditional expectation of a customer's purchase. To simplify the model, we assume that the second period is of equal length to the first (observed) period. We first derive the more generic case when only x_i is observed in the first period

$$\mathrm{E}[N^* | x_{-1}, x_0, x_1] = \sum_{i=-1}^{1} \frac{(r_i + x_i) \times \; {}_2F_1\left(r_i + x_i + 1, b_i; a_i + b_i + x_i; \frac{1}{\alpha_i+1}\right)}{(\alpha_i+1) \times \; {}_2F_1\left(r_i + x_i, b_i; a_i + b_i + x_i; \frac{1}{\alpha_i+1}\right)}, \tag{9}$$

where N^* denote the second period purchases (of a user). Then we compute the expected purchases given both the review profile x_i and n of a customer in the first period.

$$\mathrm{E}[N^*|x_{-1}, x_0, x_1, n]$$

$$= \sum_{i=-1}^{1} \frac{\sum_{l=0}^{n-\sum_{m=-1}^{1} x_m} \frac{B(a_i+x_i, l+b_i)\Gamma(l+x_i+r_i+1)}{l!(1+\alpha_i)^{l+1}} P\left(N_j+N_k=n-x_i-l \middle| X_j=x_j, X_k=x_k\right)}{\mathrm{B}(a_i+x_i, b_i)\Gamma(r_i+x_i)\ {}_2F_1\left(r_i+x_i, b_i; a_i+b_i+x_i; \frac{1}{\alpha_i+1}\right) P(N=n|X_{-1}=x_{-1}, X_0=x_0, X_1=x_1)} \qquad (10)$$

The detailed derivations of Equations (9) and (10) are presented in the Appendix. These two equations enable us to evaluate the out-of-sample performance of our proposed model.

5. MODEL EVALUATION RESULTS

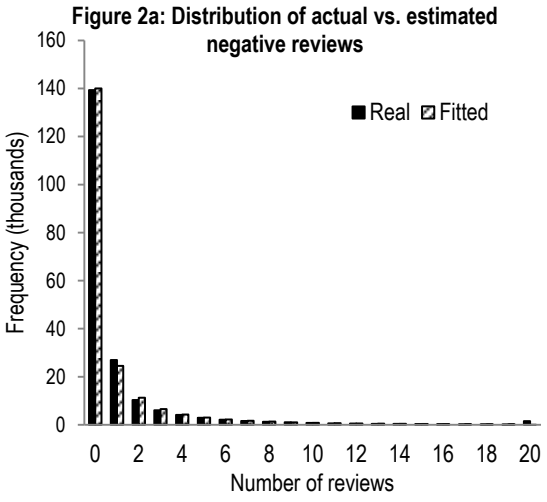
Using the Blockbuster data set described in Section 3, we estimated the parameters of the distribution as specified in Equation (3) using the maximum likelihood estimation (MLE) method. The results are presented in Table 2. Note that the estimation procedure only takes users' review data as inputs. We then evaluate the performance of this model against the actual rental data from Blockbuster using Equation (6).

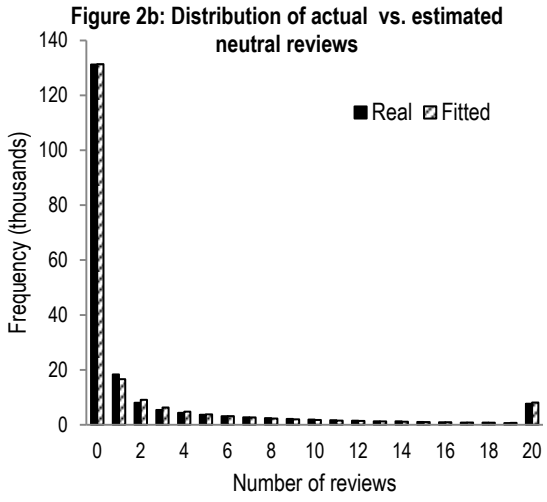
Table 2: Parameters estimated

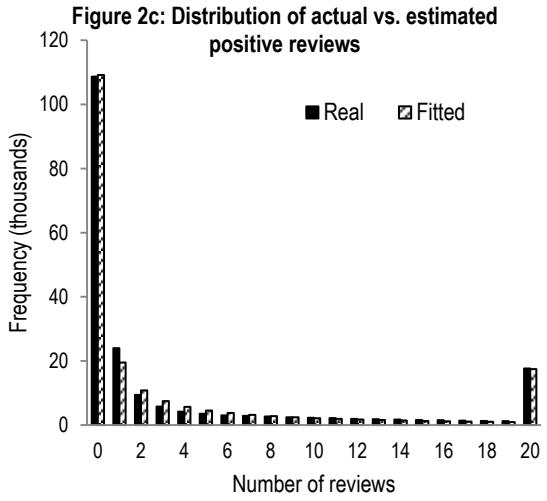
	Negative	Neutral	Positive
r_i	0.415	0.343	0.297
α_i	0.044	0.051	0.020
a_i	0.450	0.111	0.297
b_i	3.182	0.143	0.414
LL	-250,097	-308,600	-424,000

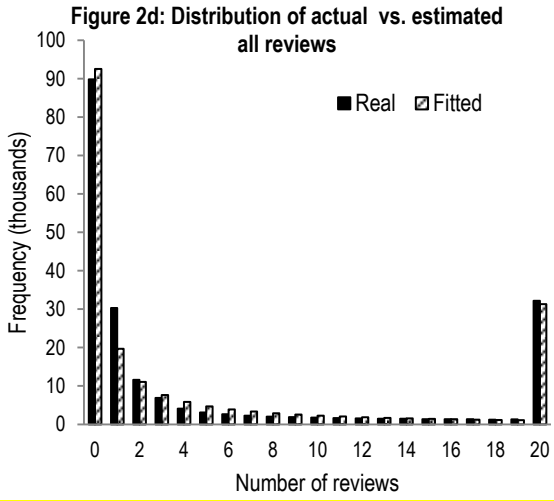
5.1. Fitness of the Model

We plot the fitness of the model for the three levels of sentiments separately. In Figure 2a, we group customers by the number of negative reviews (x_{-1}) they made. The x-axis represents the total number of (negative) reviews and the y-axis bins the number of customers in each category. The black bars (the left one in each pair of bars) are the histogram of the actual negative reviews. We then fit this underlying distribution behind the histogram using Equation (3). The shadowed bars represent the fitted (expected) results. We can visually verify that the BB/NBD model fits the actual distribution well. Formally, we compute the Theil's U index (Theil 1966), where a U-value less than 0.1 indicates good fit. The U index for our model is 0.019, suggesting satisfactory fit. In Figures 2b and 2c, we follow the same process for neutral and positive reviews (x_0, x_1) respectively. In both cases, our model fits well, with U indexes equal to 0.017 and 0.046, respectively.





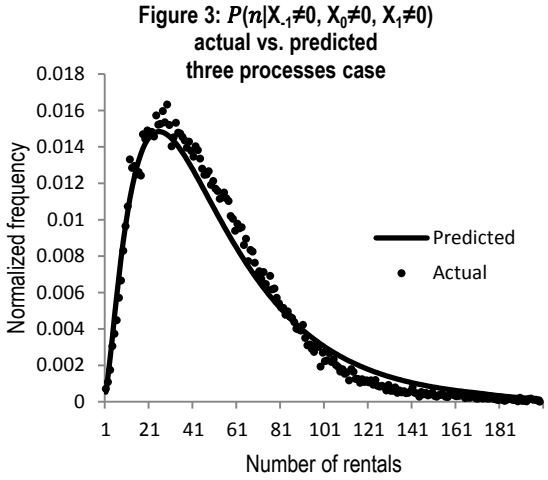




To mimic the norm of using the aggregated UGC measures in the literature, we pool the three processes into one, disregarding the existence of three different types of sentiments. This case is equivalent to assuming that customers have the same reporting probability for all their rentals regardless of their sentiments. We re-estimated the BB/NBD model on such aggregated data. The fitness turns out to be much worse as exhibited in Figure 2d, with an unsatisfactory Theil's U = 0.113. This partially demonstrates the existence of aggregation bias when the reviews of all customers (regardless of sentiments) are aggregated indiscriminately. The result disproves the hidden assumption of aggregate measures that a user's reporting probability is independent of her sentiment level. Mathematically, when we pool all the reviews together as reflected in Figure 2d, the total number of reviews X, $X = X_{-1} + X_0 + X_1$, no longer follows a BB/NBD distribution (Morrison and Schmittlein 1988).

5.2. Model Evaluation

In Figure 3, we plot the expected total number of purchases, given the observed reviews, using the evaluation scheme discussed in Section 4.5. This figure depicts the expected distribution vs. actual distribution of $P(n|X_{-1} \neq 0, X_0 \neq 0, X_1 \neq 0)$ (see Equation (8)). Theil's U index in this case is 0.099, indicating a reasonably good fit, considering the complexity of the problem and the large amount of missing data as a result of the silent customers.



To evaluate the model's predictive performance, we split the data into two periods: the first 2.5 months as the in-sample and the other 2.5 months as the out-of-sample. We use the *root mean square error (RMSE)* as the metric to measure the difference between the predicted and the actual purchases. We computed *RMSE* under three scenarios: 1) $E(N^*|X)$ representing the expected future purchase (of a customer) given her total number of reviews in the first period using the conventional BB/NBD model; 2) $E(N^*|x_i)$ representing the case when we know the review profile (x_{-1}, x_0, x_1) in the first period as computed by Equation (9); and 3) $E(N^*|x_i, n)$ representing the case when we know the customers review profile as well as the total number of purchases in the first period as shown in Equation (10).

The *RMSE* values for these three scenarios are 36.23, 18.89 and 9.47 respectively. This shows superior predictability of our proposed model over the usual BB/NBD that disregards the aggregation bias. Moreover, as expected, knowing the first period purchases greatly improves our model prediction.

5.3. Users' Reporting Probabilities

In Figure 4a, we plot the distributions of the reporting probabilities of the three different sentiment levels. The expected reporting probability is 0.06 when customers develop a negative sentiment towards a movie. This probability increases to 0.23 for the neutral ones and 0.32 for the positive ones. The differences in reporting probability at the three sentiment levels are alarmingly large. If we take an

average customer (see Table 1) as an illustration, this customer would have posted 5.59 positive reviews, 3.00 neutral ones and 1.23 negative ones, which means 65% of the times the user is satisfied with the movies and 14% of the times she is not. However, if we weight each value by its corresponding reporting probability, then we find that the average user has watched 17.5 good movies, 13.0 neutral movies and 20.5 bad movies! That is, 40% of time she encountered an unhappy experience and only 34% times she had a pleasant one. The pattern of polarity has been completely reversed!

A stronger result is shown in Figure 4b using the complimentary cumulative distribution function (CCDF) of three reporting probabilities, which is defined by one minus the cumulative distribution function (CDF). User's reporting probability when she is positive first order stochastically dominates the probability when she is negative.

Figure 4a: Density of reporting propability p_i

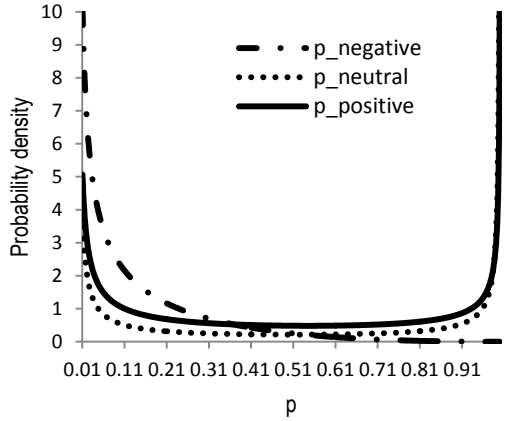
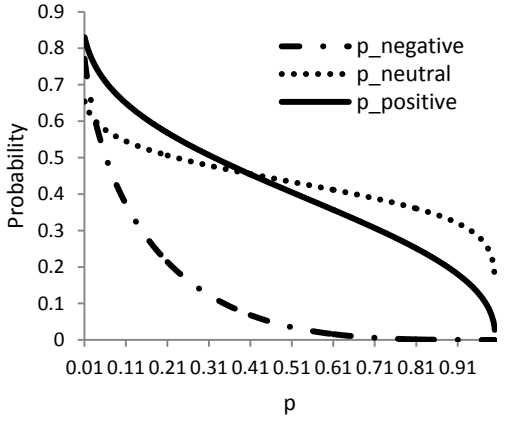


Figure 4b: Reporting propability CCDF(p_i)



6. COMPARATIVE ECONOMETRICS ANALYSES WITH VS. WITHOUT DE-BIASED UGC MEASURES

The previous two sections presented and validated our approach to estimating users' reporting probabilities. In this section, we propose a novel procedure - inverse probability weighting (IPW) - that utilizes the estimated probability to de-bias raw UGC measures.

6.1. The Inverse Probability Weighting Approach

After deriving the distribution of reporting probability, we can then use it to adjust (weight) the common aggregate UGC measures such as volume and sentiment. This "weighting" process bears similarity to the stratified sampling procedure (Wooldridge 2010, p.853) with a key difference. In stratified sampling, samples are drawn from different strata of the population with *known* probabilities; while in our study, such (reporting) probabilities of users at a particular sentiment level are *unknown* and have to be estimated first. Nevertheless, in a sense, our approach effectively stratifies a user along two

dimensions: her sentiment (towards the movie) and the number of her historical reviews at a particular sentiment level[13]. Thus an online review can be viewed as a sample drawn from a stratum according to certain "sampling" probability distribution, such as the BB/NBD distribution specified in Equation (4).

Based on Equation (4), the estimation of the expected "sampling" probability could be derived as:

$$\mathrm{E}(p_i|x_i) = \frac{(a_i + x_i)}{a_i + b_i + x_i} \frac{{}_2F_1(r_i + x_i, b_i; a_i + b_i + x_i + 1; 1/(\alpha_i + 1))}{{}_2F_1(r_i + x_i, b_i; a_i + b_i + x_i; 1/(\alpha_i + 1))}, \qquad i = -1, 0, 1. \tag{11}$$

Following Wooldridge (2010, p. 821), we then implement the IPW procedure as follows:

1. Extract all reviewers' previous reviews, and use these as inputs to estimate the model parameters in Equation (3).
2. For each focal movie, identify this movie's reviewers and all of their historical reviews.
3. Calculate each reviewer's expected reporting probability p_i for the focal movie using Equation (11), and weight her UGC measures (e.g., sentiment and volume) using the IPW scheme, i.e., assigning a weight of $1/\mathrm{E}(p_i|x_i)$ to the review.
4. Derive the de-biased UGC measures by aggregating over all the reviews of this movie.

6.2. The Comparative Analysis

A common theme in online review studies is to investigate the role of reviews on sales. Such studies often use sales as a dependent variable and various measures of online reviews as independent variables. We validate our approach by replicating the seminal UGC study of Forman et al. (2008) which establishes a relationship between Amazon book sales rank and volume, valence (average rating) and reviewer identity (see Equation (2) of Forman et al. 2008). We compare such a model's performance *with* versus *without* de-biased measures. The canonical econometrics model using volume and sentiment to model movie rental at Blockbuster can be expressed as follows:

$$\begin{aligned}\mathrm{Log}\left(SalesRank_{i,t}\right) = \alpha + \beta_1 \mathrm{Log}\left(NumberofReviews_{i,t-1}\right) + \beta_2 AvgReviewRating_{i,t-1} \\ +\beta_3 \mathrm{Log}\left(WeekElapsed_{i,t}\right) + \mu_i + \varepsilon_{i,t}\,.\end{aligned} \tag{12}$$

This represents a simple fixed effects model that closely emulates Equation (2) of Forman et al. (2008), with two key adjustments. First, unlike Amazon, at Blockbuster there is no mechanism for reviewers to disclose their identities and hence the key variable in Forman et al. (2008), reviewer identity, is not replicated here. Second, product price is used as a control variable in their model. However, in our case, since rental price is uniform across DVDs for all subscribers within the same subscription type (see footnote 9), there is no need to control for price. All other specifications are the same as Forman et al. (2008). The notations and definitions of each variable are summarized in Table 3.

[13]For example, all customers who are negative toward the focal movie, and already have five negative reviews before watching this movie can be perceived to be in the same stratum.

Among the Blockbuster new-releases from October 2009 to February 2010, we focus on 53 movies that have at least 6 weeks of life span[14]. We then run two versions of regression models as specified by Equation (12), one with our de-biased measures and the other with unadjusted ones, as commonly used in the UGC literature. The results are presented in Table 4.

Table 3: Variables

Variable	Description
i	Index for movie
t	Index for time
$SalesRank_{i,t}$	The rank of rental sales of movie i in week t
$NumberofReviews_{i,t-1}$	Total number of reviews posted online for movie i in week $t-1$
$AvgReviewRating_{i,t-1}$	Average star rating of users for movie i in week $t-1$
$WeekElapsedk_{i,t}$	Number of weeks after the release of movie i
μ_i	Fixed effects for movie i

Table 4: Regression results

Variable	Model without adjustment	Model with adjustment
$Log\_NumberofReviews_{i,t-1}$	-0.224**	
$AvgReviewRating_{i,t-1}$	-0.351	
$Log\_Adjust\_NumberofReviews_{i,t-1}$		-0.184*
$Adjust\_AvgReviewRating_{i,t-1}$		-1.196**
$WeekElapsedk_{i,t}$	1.611***	1.549***
R^2	0.961	0.962
N	53	53

**** p<0.0001, *** p<0.001, "Log" indicates a logarithmic transformation. "Adjust" means our adjusted measures.

First, for volume, both models indicate a significant and positive impact (see coefficients for *Log_NumberofReviews* and *Log_Adjust_NumberofReviews*)[15]. However, the results for sentiment yield some interesting differences. According to our model, *Adjust_AvgReviewRating* is found to positively and significantly impact sales. In the unadjusted model, *AvgReviewRating* does not turn out to be significant, consistent with the finding in Forman et al. (2008). As we have elaborated, these measures are potentially subject to reporting bias which may lead to biased estimates. We believe that the reporting bias is one of the main causes for the mixed findings on sentiment in the UGC literature. Such conflicting findings can

[14]Blockbuster defines the life span of a movie as having at least 100 rentals per day (or 700 rentals per week). Rental movies usually have very short life span (less than 2 months). We examine the first 6 weeks' rentals of a movie.
[15] Our dependent variable is the sales rank. A negative coefficient indicates a positive impact on sales.

result in erroneous managerial decisions. By introducing IPW, one of our objectives is to reconcile such inconsistent results.

Second, note that the R^2 values are comparable in both models. We would like to point out that R^2 is not an appropriate evaluation criterion to compare the two models here though. In fact, R^2 can be either improved or decreased after applying the de-biasing procedure. This resembles the two-stage least square (2SLS) estimation when instrument variables (IV) are used to address endogeneity. It is not guaranteed that the 2SLS procedure will necessarily result in higher R^2 than the biased OLS in the presence of endogeneity. But it is still necessary to apply IV estimation to account for the biases introduced by endogenous variables. Likewise, the goal of IPW is to ensure unbiased estimation, not necessarily better fit, as held by Başeret al. (2006).

As we have demonstrated, reviewers' selective reporting behaviors can cause a systematic bias that may not necessarily be random or two-sided as is assumed in the treatment of Error-in-Variable scenarios (Wooldridge 2010, p. 78). Blockbuster's review system exhibits a one-sided bias towards the positive side. Wooldridge (2010) shows that the (random) error in variable causes attenuation (i.e., smaller estimates) of the parameters. In other words, after de-biasing, the parameter estimates should become larger. This is evident in our estimation of sentiment: -1.196 (de-biased) vs. -0.351 (original). However, the coefficients of volume turn smaller after adjustment, appearing to be at odds with the prediction of the attenuation bias theory. Nevertheless, a more careful analysis below shows that this is to be expected. This difference is mainly due to the fact that the unadjusted volume is smaller than the adjusted one, leading to inflated magnitude of estimates. This could be further illustrated as follows:

Let $\log(V^*) = \log(V) \cdot p_\varepsilon$, where V^* is the unadjusted review volume and V is the adjusted volume. Without loss of generality, we model the error p_ε in a multiplicative manner to facilitate derivation. According to our proposed approach, $0 < p_\varepsilon < 1$. Then,

$$|\beta_1^*| = \left|\frac{\partial \log(Sales)}{\partial \log(V^*)}\right| = \left|\frac{\partial \log(Sales)}{\partial \log(V)}\frac{\partial \log(V)}{\partial \log(V^*)}\right| = \left|\frac{\partial \log(Sales)}{\partial \log(V)}\frac{1}{p_\varepsilon}\right| = \left|\beta_1 \frac{1}{p_\varepsilon}\right| > |\beta_1|. \quad (13a)$$

Similarly, for sentiment, let $S^* = S \cdot q_\varepsilon$, where S^* is the unadjusted average rating and S is the adjusted average rating. According to our approach, the adjusted rating is usually smaller than the unadjusted one. Since q_ε is in general greater than one, we have:

$$|\beta_2^*| = \left|\frac{\partial \log(Sales)}{\partial S^*}\right| = \left|\frac{\partial \log(Sales)}{\partial S}\frac{\partial S}{\partial S^*}\right| = \left|\frac{\partial \log(Sales)}{\partial S}\frac{1}{q_\varepsilon}\right| = \left|\beta_2 \frac{1}{q_\varepsilon}\right| < |\beta_2|. \quad (13b)$$

As Table 4 shows, our results are exactly as Equation (13a) and (13b) predict: 0.224>0.184 for volume and 0.351<1.196 for sentiment.

7. DISCUSSIONS

In this paper, we examine two biases commonly plaguing UGC studies, the aggregation and reporting biases. We propose a method that simultaneously addresses them by estimating a user's tendency to produce UGC at different levels of user sentiments. The model is calibrated on a real dataset obtained from Blockbuster. The results clearly demonstrate the existence of the reporting and aggregation biases embedded within Blockbuster's review system. By accounting for these two biases, we are able to infer the actual total sales (movie rentals) accurately using only the observed review data, which shows the promise of this approach. Further, we develop an IPW approach to de-bias several common UGC measures including volume and sentiment. Our econometrics analysis by replicating a seminal UGC paper demonstrates the ability of our approach to reconcile the conflicting findings in UGC studies via addressing these two biases.

Our results suggest that managers should account for these prevalent biases in UGC. For example, the average rating of the movies in Blockbuster dataset was 3.89. After addressing these two biases, the new (and more believable) average rating was reduced to 2.88. This shifts an average customer's sentiment from positive to negative (or close to neutral). Incognizant of this difference may mislead a myopic manager to make erroneous managerial decisions based on biased UGC samples.

Our de-biasing approach (IPW) offers a direct solution for online review hosts, such as Yelp.com or Epinions.com. As the debates between YouTube.com and Yelp.com at the beginning of the paper show, the inflated ratings affect the credibility and thus usability of online reviews, eventually dampening readers' trust on the entire review system. Yelp's simple solution (i.e., factoring in the experience of reviewer and the helpfulness of reviews) essentially assumes only experienced or elite reviewers are trustworthy. This may cause unwanted side effects such as discouraging less experienced reviewers to participate. Our proposed adjustment approach, however, considers every reviewer's opinion by utilizing her historical behaviors.

Our approach is broadly applicable to a wide range of online retailers, such as Amazon, eBay, or other online retailers as long as they keep track of their customer's review history, which they normally do. The de-biased measures derived from our method can serve as valuable inputs for retailers to improve their CRM and recommendation systems. An interesting future research question is how our de-biasing procedure improves the accuracy of recommender systems.

Our work is not without limitations. Our approach assumes that historical reviews of a reviewer are available. However, sufficient amount of historical reviews of users may be needed to effectively apply our approach. In our experiments, we show that, with just 5-month of review data, we are able to recover users' reporting probability well. Another limitation is that we do not directly observe each customer's decision making process in producing UGC. We thus rely on a series of assumptions under the BB/NBD

framework to model this decision process, in an endeavor to uncover the silent users' behaviors. A potential straightforward approach is to survey a sample of silent users directly on their sentiments. A procedure known as "list augmentation" or "database augmentation" (Crosby et al 2002, Kamakura and Wedel 2003) can be then applied to combine this survey sample data with what are available to researchers. Interested readers are referred to Du et al. (2007) for more details of such a procedure.

As a concluding remark, one may argue that the unadjusted raw UGC (e.g., the reviews displayed on Blockbuster.com) may be of interest in its own right, as these reviews are what customers actually see before making purchasing decisions. Thus, it can be argued that it would be useful to build a regression model of movie rentals on the raw review measures. We would like to counter however, that, though it may be true that unsophisticated customers may rely on such biased review measures, a better informed customer can see through such biased representation of reviews and form her own "model" to adjust the reviews accordingly. In a sense, what we are trying to accomplish is to more accurately emulate a customer's decision model to de-bias such unrepresentative measures**.**

References

Abhishek, V., K. Hosanagar, P. Fader. 2011. On aggregation bias in sponsored search data: Existence and implications. Working paper, University of Pennsylvania, Philadelphia, PA.

Abramowitz, M. and I. A. Stegun. 1965. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover publications, New York.

Allen, S. G. 1981. An empirical model of work attendance. *Rev. Econ. Statist.* **63**(1), 77-87.

Barmby, T. A., C. D. Orme, J. G. Treble. 1991. Worker absenteeism: an analysis using microdata. *Econ. J.* **101**(405), 214-229.

Basuroy, S., S. Chatterjee, S. A. Ravid. 2003. How critical are critical reviews? The box office effects of film critics, star power, and budgets. *J. Marketing.* **67**(4), 103-117.

Başer, O., J. C. Gardiner, C.J. Bradley, H. Yüce, C. Given. 2006. Longitudinal analysis of censored medical cost data. *Health Econ.* **15**(5), 513–525.

Bickart, B., D. Schmittlein. 1999. The distribution of survey contact and participation in the United States: Constructing a survey-based estimate. *J. Marketing Res.* **36**(2), 286-294.

Bickel, P. J., E. A. Hammel, J. W. O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science.* **187**(4175), 398-404.

Chatfield, C., G. J. Goodhardt. 1970. The Beta-Binomial model for consumer purchasing behavior. *J. Roy. Statist. Soc. Ser. C.* **19**(3), 240-250.

Chen, P., S. Dhanasobhon, M. Smith. 2008. All reviews are not created equal: The disaggregate impact of reviews and reviewers at Amazon.com. Working paper, Temple University, Philadelphia, PA.

Chevalier, J. A., D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* **43**(3), 345-354.

Clemons, E. K., G. G. Gao, L. M. Hitt. 2006. When online reviews meet hyper differentiation: A study of the craft beer industry. *J. Management Inform. Systems.* **23**(2), 149-171.

Crosby, A., S. Johnson, R. Quinn. 2002. Is survey research dead? *Marketing Management.* **11**(3), 24-29.

Du, R., W. Kamakura, C. Mena. 2007. Size and share of customer wallet. *J. Marketing.* **71**(4), 94-113.

Dellarocas, C., C. A. Wood. 2008. The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Sci.* **54**(3), 460-476.

Duan, W., B. Gu, A. B. Whinston. 2008a. Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems.* **45**(4), 1007-1016.

Duan, W., B. Gu, A. B. Whinston. 2008b. The dynamics of online word-of-mouth and product sales: An empirical investigation of the movie industry. *J. Retailing.* **84**(2), 233-242.

Dunn, R., S. Reader, N. Wrigley. 1983. An investigation of the assumptions of the NBD model as applied to purchasing at individual stores. *J. Roy. Statist. Soc. Ser. C.* **32**(3), 249-259.

Etzion, H., N. Awad. 2007. Pump up the volume? Examining the relationship between number of online reviews and sales: is more necessarily better? *ICIS 2007 Proceedings.* Paper 120. (available at http://aisel.aisnet.org/icis2007/120)

Fader, P. S., B. G. S. Hardie. 2000. A note on modeling underreported Poisson counts. *J. Appl. Statist.* **27**(8), 953-964.

Fader, P. S., B. G. S. Hardie. 2010. Customer-base valuation in a contractual setting: The perils of ignoring heterogeneity. *Marketing Sci.* **29**(1), 85-93.

Fader, P. S. and D. C. Schmittlein. 1993. Excess behavioral loyalty for high-share brands: Deviations from the Dirichletmodel for repeat purchasing. *J. Marketing Res.* **30**(4), 478-493.

Forman, C., A. Ghose, B. Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* **19**(3), 291-313.

Ghose, A., P. G. Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowledge Data Engrg.* **23**(10), 1498-1512.

Goodhardt, G. J., A. S. C. Ehrenberg, C. Chatfield. 1984. The Dirichlet: a comprehensive model of buying behavior. *J. Roy. Statist. Soc. Ser. A.* **147**(5), 621-655.

Greene, J. D. 1982. *Consumer behavior models for non-statisticians*. Praeger Publishers, New York.

Hausman, J., D. McFadden. 1984. Specification tests for the multinomial logit model. *Econometrica.* **52**(5), 1219-1240.

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

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

Jabr, W., Z.Zheng. 2011. The competition effect of WOM and recommendations on customer choice: An empirical investigation. Working paper, University of Texas at Dallas, Richardson, TX.

Jeuland, A. P., F. M. Bass, G. P. Wright. 1980. A multibrand stochastic model compounding heterogeneous Erlang timing and multinomial choice processes. *Oper. Res.* **28**(2), 255-277.

Johnson, L., S. Kotz, A. Kemp. 1992. *Univariate Discrete Distributions*, 2nd ed. John Wiley & Sons, New York.

Kamakura, W. A., M. Wedel. 2003. List augmentation with model based multiple imputation: A case study using a mixed outcome factor model. *Statistica Neerlandica*. **57**(1), 46-57.

Kelejian, H. H. 1995. Aggregated heterogeneous dependent data and the logit model: A suggested approach. *Econ. Letters.* **47**(3-4), 243-248.

Li, X., L. M. Hitt. 2008. Self-selection and information role of online product reviews. *Inform. Systems Res.* **19**(4), 456-474.

Liu, Y. 2006. Word of mouth for movies: Its dynamics and impact on box office revenue. *J. Marketing.* **70**(3), 74-89.

Morrison, D. G., D. C. Schmittlein. 1981. Predicting future random events based on past performance. *Management Sci.* **27**(9), 1006-1023.

Morrison, D. G., D. C. Schmittlein. 1988. Generalizing the NBD model for customer purchases: What are the implications and is it worth the effort? *J. Bus. Econ. Statist.* **6**(2), 145-159.

Mudambi, S. M., D. Schuff. 2010. What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quart.* **34**(1), 185-200.

Ross, S. M. 2003. *Introduction to Probability Models*, 8th ed. Academic Press, San Diego, CA.

Schmittlein, D. C., A. C. Bemmaor, D. G. Morrison. 1985. Why does the NBD model work? Robustness in representing product purchases, brand purchases and imperfectly recorded purchases. *Marketing Sci.* **4**(3), 255-266.

Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B.* **13**(2), 238-241.

Theil, H. 1966. *Applied Economic Forecasting*. North-HoIland Publishing, Amsterdam.

Winkelmann, R. 1996. Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Econ.* **21**(4), 575-587.

Winkelmann, R. 2008. *Econometric Analysis of Count Data*, 5th ed. Springer, Berlin.

Wooldridge, J. 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Cambridge, MA.

Zheng, Z., P. Fader, B.Padmanabhan. 2011. From business intelligence to competitive intelligence: Inferring competitive measures using augmented site-centric data. *Inform. Systems Res*. Forthcoming.

Appendix

1. Derivation of reporting probability conditional on observed x_i and n

We derive the reporting probability (of a random user) given her review profile (x_{-1}, x_0, x_1) and total purchase of the product n as follows:

$$g(p_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n)=\int_0^\infty h(\lambda_i,p_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n)d\lambda_i,$$

where $h(\lambda_i,p_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n)$is the conditional joint density of λ_i and p_i. Note that

$$h(\lambda_i,p_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n)$$

$$=\frac{P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,\lambda_i,p_i)P(X_{-1}=x_{-1},X_0=x_0,X_1=x_1|\lambda_i,p_i)f(\lambda_i)g(p_i)}{P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)P(X_{-1}=x_{-1},X_0=x_0,X_1=x_1)}$$

$$=\frac{P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,\lambda_i,p_i)P(X_i=x_i|\lambda_i,p_i)f(\lambda_i)g(p_i)}{P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)P(X_i=x_i)}$$

$$=\frac{(\lambda_i p_i)^{x_i}e^{-\lambda_i p_i}}{x_i!}\frac{{\alpha_i}^{r_i}{\lambda_i}^{r_i-1}e^{-\lambda_i\alpha_i}}{\Gamma(r_i)}\frac{{p_i}^{a_i-1}(1-p_i)^{b_i-1}}{\mathrm{B}(a_i,b_i)}$$

$$\times\frac{\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\left\{\frac{[\lambda_i(1-p_i)]^l e^{-\lambda_i(1-p_i)}}{l!}P\left(N_j+N_k=n-x_i-l\middle|X_j=x_j,X_k=x_k\right)\right\}}{P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)P(X_i=x_i)}$$

$$=\frac{{\alpha_i}^{r_i}p_i^{a_i+x_i-1}\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\left\{\frac{(1-p_i)^{l+b_i-1}{\lambda_i}^{l+x_i+r_i-1}e^{-\lambda_i(1+\alpha_i)}}{l!}P\left(N_j+N_k=n-x_i-l\middle|X_j=x_j,X_k=x_k\right)\right\}}{x_i!\Gamma(r_i)\mathrm{B}(a_i,b_i)P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)P(X_i=x_i)},\qquad(A-1)$$

where $j,k\in\{-1,0,1\}\ and\ j\neq i\neq k$ denote the other two sentiment levels that are different from i. $P\left(N_j+N_k=n-x_i-l\middle|X_j=x_j,X_k=x_k\right)$ and $P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)$ can be computed via the convolution of $P(N_i=n_i|X_i=x_i)$. We also know that

$$P(N_i=n_i|X_i=x_i)$$

$$=\frac{\Gamma(r_i+n_i)}{\Gamma(r_i+x_i)(n_i-x_i)!}\left(\frac{1}{\alpha_i+1}\right)^{n_i-x_i}\frac{\Gamma(a_i+b_i+x_i)}{\Gamma(a_i+b_i+n_i)}\frac{\Gamma(b_i+n_i-x_i)}{\Gamma(b_i)}$$

$$/\ {}_2F_1\left(r_i+x_i,b_i;a_i+b_i+x_i;\frac{1}{\alpha_i+1}\right),\quad i=-1,0,1.$$

It follows that

$$g(p_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n)$$

$$=\int_0^\infty \frac{{\alpha_i}^{r_i}{p_i}^{a_i+x_i-1}\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\left\{\frac{(1-p_i)^{l+b_i-1}\lambda_i{}^{l+x_i+r_i-1}e^{-\lambda_i(1+\alpha_i)}}{l!}P\left(N_j+N_k=n-x_i-l\middle|X_j=x_j,X_k=x_k\right)\right\}}{x_i!\,\Gamma(r_i)\mathrm{B}(a_i,b_i)P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)P(X_i=x_i)}\,d\,\lambda_i$$

$$=\frac{{p_i}^{a_i+x_i-1}\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\left\{\frac{(1-p_i)^{l+b_i-1}\Gamma(l+x_i+r_i)}{l!(\alpha_i+1)^l}P\left(N_j+N_k=n-x_i-l\middle|X_j=x_j,X_k=x_k\right)\right\}}{\mathrm{B}(a_i+x_i,b_i)\Gamma(r_i+x_i)\;{}_2F_1\left(r_i+x_i,b_i;a_i+b_i+x_i;\frac{1}{\alpha_i+1}\right)P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)}.\quad(A-2)$$

$$\mathrm{E}[p_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n]$$

$$=\int_0^\infty p_i*\frac{{p_i}^{a_i+x_i-1}\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\left\{\frac{(1-p_i)^{l+b_i-1}\Gamma(l+x_i+r_i)}{l!(\alpha_i+1)^l}P\left(N_j+N_k=n-x_i-l\middle|X_j=x_j,X_k=x_k\right)\right\}}{\mathrm{B}(a_i+x_i,b_i)\Gamma(r_i+x_i)\;{}_2F_1\left(r_i+x_i,b_i;a_i+b_i+x_i;\frac{1}{\alpha_i+1}\right)P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)}\,d\,p_i$$

$$=\int_0^\infty \frac{\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\left\{\frac{{p_i}^{a_i+x_i}(1-p_i)^{l+b_i-1}\Gamma(l+x_i+r_i)}{l!(\alpha_i+1)^l}P\left(N_j+N_k=n-x_i-l\middle|X_j=x_j,X_k=x_k\right)\right\}}{\mathrm{B}(a_i+x_i,b_i)\Gamma(r_i+x_i)\;{}_2F_1\left(r_i+x_i,b_i;a_i+b_i+x_i;\frac{1}{\alpha_i+1}\right)P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)}\,d\,p_i$$

$$=\frac{\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\left\{\frac{\mathrm{B}(a_i+x_i+1,l+b_i)\Gamma(l+x_i+r_i)}{l!(\alpha_i+1)^l}P\left(N_j+N_k=n-x_i-l\middle|X_j=x_j,X_k=x_k\right)\right\}}{\mathrm{B}(a_i+x_i,b_i)\Gamma(r_i+x_i)\;{}_2F_1\left(r_i+x_i,b_i;a_i+b_i+x_i;\frac{1}{\alpha_i+1}\right)P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)}.\quad(\mathrm{A}-3)$$

2. Predicting the total number of purchases in the second (equal length) period after observing x_i and n at first period

$$\mathrm{E}[N^*|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n]$$

$$=\mathrm{E}[N_{-1}^*+N_0^*+N_1^*|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n]$$

$$=\sum_{i=-1}^{1}\mathrm{E}[N_i^*|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n]$$

$$=\sum_{i=-1}^{1}\mathrm{E}[\lambda_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n]$$

$$=\sum_{i=-1}^{1}\int_0^\infty \lambda_i f[\lambda_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n]\,d\lambda_i$$

Where

$$f[\lambda_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n]=\int_0^1 h(\lambda_i,p_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n)dp_i,$$

and $h(\lambda_i,p_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n)$ has been derived in Equation (A-1).

Thus,

$$f[\lambda_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n]$$

$$=\int_0^1 h(\lambda_i,p_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n)dp_i$$

$$=\frac{{\alpha_i}^{r_i}\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\left\{\int_0^1\frac{p_i{}^{a_i+x_i-1}(1-p_i)^{l+b_i-1}{\lambda_i}^{l+x_i+r_i-1}e^{-\lambda_i(1+\alpha_i)}}{l!}P(N_j+N_k=n-x_i-l|X_j=x_j,X_k=x_k)dp_i\right\}}{x_i!\Gamma(r_i)\mathrm{B}(a_i,b_i)P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)P(X_i=x_i)}$$

$$=\frac{{\alpha_i}^{r_i}\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\frac{B(a_i+x_i,l+b_i){\lambda_i}^{l+x_i+r_i-1}e^{-\lambda_i(1+\alpha_i)}}{l!}P(N_j+N_k=n-x_i-l|X_j=x_j,X_k=x_k)}{x_i!\Gamma(r_i)\mathrm{B}(a_i,b_i)P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)P(X_i=x_i)}. \quad (\mathrm{A}-4)$$

Finally

$$\mathrm{E}[N^*|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n]=\sum_{i=-1}^{1}\int_0^\infty \lambda_i f[\lambda_i|X_{-1}=x_{-1},X_0=x_0,X_1=x_1,N=n]\,d\lambda_i$$

$$=\sum_{i=-1}^{1}\int_0^\infty\frac{{\alpha_i}^{r_i}\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\frac{B(a_i+x_i,l+b_i){\lambda_i}^{l+x_i+r_i}e^{-\lambda_i(1+\alpha_i)}}{l!}P\left(N_j+N_k=n-x_i-l\middle|X_j=x_j,X_k=x_k\right)}{x_i!\,\Gamma(r_i)\mathrm{B}(a_i,b_i)P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)P(X_i=x_i)}d\lambda_i$$

$$=\sum_{i=-1}^{1}\frac{{\alpha_i}^{r_i}\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\frac{B(a_i+x_i,l+b_i)\Gamma(l+x_i+r_i+1)}{l!(1+\alpha_i)^{l+x_i+r_i+1}}P\left(N_j+N_k=n-x_i-l\middle|X_j=x_j,X_k=x_k\right)}{x_i!\,\Gamma(r_i)\mathrm{B}(a_i,b_i)P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)P(X_i=x_i)}$$

$$=\sum_{i=-1}^{1}\frac{\sum_{l=0}^{n-\sum_{m=-1}^{1}x_m}\frac{B(a_i+x_i,l+b_i)\Gamma(l+x_i+r_i+1)}{l!(1+\alpha_i)^{l+1}}P\left(N_j+N_k=n-x_i-l\middle|X_j=x_j,X_k=x_k\right)}{\mathrm{B}(a_i+x_i,b_i)\Gamma(r_i+x_i)\;{}_2F_1\left(r_i+x_i,b_i;a_i+b_i+x_i;\frac{1}{\alpha_i+1}\right)P(N=n|X_{-1}=x_{-1},X_0=x_0,X_1=x_1)}.$$

$$(A-5)$$

The simpler case when only x_i is known is as:

$$\mathrm{E}[N^*|X_{-1}=x_{-1},X_0=x_0,X_1=x_1]=\sum_{i=-1}^{1}E[N_i^*|X_i=x_i]=\sum_{i=-1}^{1}E[\lambda_i|X_i=x_i]$$

$$=\sum_{i=-1}^{1}\frac{(r_i+x_i)\times\;{}_2F_1\left(r_i+x_i+1,b_i;a_i+b_i+x_i;\frac{1}{\alpha_i+1}\right)}{(\alpha_i+1)\times\;{}_2F_1\left(r_i+x_i,b_i;a_i+b_i+x_i;\frac{1}{\alpha_i+1}\right)}. \qquad (A-6)$$