2-2011

# Fraud detection in online consumer reviews

Nan HU

Ling LIU

Vallbh SAMBAMURTHY

Published in Decision Support Systems
Volume 50, Issue 3, February 2011, Pages 614-626
https://doi.org/10.1016/j.dss.2010.08.012


# Fraud detection in online consumer reviews ☆

Nan Hu [a,b,*], Ling Liu [a], Vallabh Sambamurthy [c]

[a] Department of Accounting and Finance, University of Wisconsin Eau Claire, 105 Garfield Ave, Eau Claire, United States
[b] School of Information Systems, Singapore Management University, 80 Stamford Road, Singapore
[c] Center for Leadership of the Digital Enterprise Eli Broad, College of Business, Michigan State University, East Lansing, MI 48824-1122, United States

*Keywords:*
Online word of mouth
Manipulation
Self-selection
Price
Time-series

A B S T R A C T

Increasingly, consumers depend on social information channels, such as user-posted online reviews, to make purchase decisions. These reviews are assumed to be unbiased reflections of other consumers' experiences with the products or services. While extensively assumed, the literature has not tested the existence or non-existence of review manipulation. By using data from Amazon and Barnes & Noble, our study investigates if vendors, publishers, and writers consistently manipulate online consumer reviews. We document the existence of online review manipulation and show that the manipulation strategy of firms seems to be a monotonically decreasing function of the product's true quality or the mean consumer rating of that product. Hence, manipulation decreases the informativeness of online reviews. Furthermore though consumers understand the existence of manipulation, they can only partially correct it based on their expectation of the overall level of manipulation. Hence, vendors are able to change the final outcomes by manipulating online reviewers. In addition, we demonstrate that at the early stages, after an item is released to the Amazon market, both price and reviews serve as quality indicators. Thus, at this stage, a higher price leads to an increase in sales instead of a decrease in sales. At the late stages, price assumes its normal role, meaning a higher price leads to a decrease in sales. Finally, on average, there is a higher level of manipulation on Barnes & Noble than on Amazon.

## 1. Introduction

The rapid adoption of Web 2.0 has unleashed a wave of innovations that might change the way customers acquire information to make product purchases or stock investment decisions. The growth of Web 2.0 has enabled consumers to post reviews describing their experiences with products, product vendors, or service providers and make them available to other prospective consumers. In fact, the marketing literature suggests that consumers depend on online product reviews to make purchase decisions [3,5]. Capital markets research has revealed that the information conveyed by stock message boards are used by investors [1], and a shock to the message board postings is negatively associated with future stock returns [12].

Since consumers increasingly depend on information released through social online channels, such as consumer-generated content, to make product or services purchase decisions, the quality and truthfulness of information available to them is important. Do various entities, such as companies, vendors, publishers, or writers, actively engage in word-of-mouth manipulation, either directly or indirectly, with the goal of changing consumers' final decisions? Such practices are not new for information released through traditional information channels. For example, a rich earnings management literature has revealed that managers deliberately misrepresent financial reports in order to smooth their firm's income, meet a pre-specified target, and get better compensation.

We define review fraud as occurring when online vendors, publishers, or authors write “consumer” reviews by posing as real customers. An email interview with Jonathan Carson, CEO of BuzzMetrics, reveals that promoting new CD releases through chat promotion is almost an industry standard [11]. Such a practice exists even for highly reputable vendors, such as Amazon. In April 2004 James Marcus, a former senior editor for Amazon.com, wrote an alarming article in *The Washington Post* to discuss review fraud. Based on an analysis of reviews of just a few thousand reviewers, he found that a large number of authors on Amazon had got favorable reviews from their friends, relatives, colleagues or paid professionals. In some

☆ The authors would like to thank Ramnath Chellappa, Tridas Mukhopadhyay, and seminar participants at the National University of Singapore for their valuable feedback on earlier versions of this manuscript. The authors also thank Bin Chen for his research assistance for this study. All remaining errors and omissions are our responsibility.
* Corresponding author.
*E-mail addresses:* hun@uwec.edu, hunan@smu.edu.sg (N. Hu), liul@uwec.edu (L. Liu), sambamurthy@bus.msu.edu (V. Sambamurthy).

cases, these authors even wrote reviews for their own books.[1] Furthermore, such fraud has caused financial loss to society as well.[2]

Recent research concludes that word-of-mouth (WOM) communication is a valuable marketing resource for consumers and marketers with critical implications for a product's success. This literature provides useful insights by linking online reviews with sales. It shows a positive correlation between the average review score and product sales [4–6]. However, there is one implicit but essential assumption in this literature that researchers take for granted as being true, which is:

**Assumption 1.** Online reviews are written by actual previous customers, not publishers or vendors, etc. Therefore, online reviews reflect either the actual product quality or the product's relative true quality.

If the above assumption is true, then online reviews should reflect a products' true quality; or, all other information (e.g., price, product category, manufacturer, vendor, and shipping terms) being the same, a product with a higher mean consumer product rating should be assumed to have higher quality. This assumption is crucial in justifying the linkage between online reviews and sales. However, the existence of review fraud would invalidate such an assumption and cast doubts on the association between product quality and consumer reviews. If online reviews are indeed written by actual previous customers, then online reviews can help new customers reduce the uncertainties involved in inferring product quality, thus resulting in an increased conversation rate and higher sales. However, if online vendors, publishers, and authors are all able to write "consumer" reviews, then instead of being an uncertainty "reducer", online reviews might become an uncertainty enhancer. In such a case, consumers' beliefs about product quality and vendor reputations derived from online reviews might be totally misleading.

To date, there have been a few analytical studies investigating review fraud [2,11]. Drawing on the observation that the music industry is known to hire professional marketers to write favorable consumer opinions to promote the sales of new albums, Mayzlin [11] built an analytical game theory model in which two competing firms send anonymous messages recommending their own products. Dellarocas [2] analytically shows that if every firm's manipulation strategy monotonically increases with regard to that firm's true quality, then manipulation of online reviews increases the informativeness of online reviews. Under such a circumstance, manipulation increases the separation of the distributions of ratings and will help consumers make better purchase decisions. Even if there is manipulation, consumers are smart and can adjust their interpretation of online opinions accordingly [2]. Combining the implicit assumption stated above (Assumption 1) with these analytical works, we have the following revised assumption based on previous literature:

**Assumption 2.** Online reviews are written by actual previous customers and not publishers or vendors. Even if there is manipulation, consumers are smart and can adjust their interpretation of online opinions accordingly [2]. Further, as long as the manipulation is monotonically increasing with regard to a product's true quality (i.e., if it is more likely for higher quality vendors to engage in review manipulation), then online reviews with the existence of review fraud are even more informative than when there is no review fraud.

If consumers are indeed smart and if the manipulation is monotonically increasing with respect to (w.r.t) to product quality, then we need not worry about empirically testing manipulation of online reviews because under such a circumstance, online reviews are more informative. However, are these assumptions true?

In this paper, we analytically and empirically study temporal behaviors of online reviews and address the following research questions:

- Does review fraud actually exist? Is review manipulation a prevalent phenomenon or does it just happen occasionally?
- What types of vendors are more likely to manipulate online reviews: those selling high-quality products or those selling low-quality products? Vendors that receive higher average ratings for their products, or those with lower average ratings?
- Are consumers smart enough to filter out the manipulation as Dellarocas [2] suggests? Are they able to correct for this bias in their purchase decisions? What quality indexes do they use to make purchase decisions in view of the existence of review fraud?
- Is online review fraud a common phenomenon across different websites?

This paper proceeds as follows. Section 2 studies the mean-reverse phenomenon of consumer reviews to motivate our study. By studying the temporal patterns of online reviews, we show that there might be two potential drivers which are consumer taste difference and/or review manipulation that force rating decreases over time. As a nature follow-up question of Section 2, Section 3 answers whether a pure consumer taste difference without manipulation can be the sole underlying driving force. We conclude that we cannot rule out manipulation as one of the potential drivers. The temporal patterns of online reviews can be either driven by pure manipulation or by a joint force of consumer taste difference and manipulation. Section 4 seeks to answer the question of whether low-quality or high-quality vendors are more likely to manipulate consumer reviews. Section 5 analyzes whether consumers correct for manipulation bias when making purchase decisions. Section 6 answers how customers make purchase decisions when manipulation exists. Section 7 checks the robustness of our findings by comparing the online review manipulation between Amazon and Barnes & Nobel. Section 8 contains discussion of the findings, their implications, and some concluding remarks.

## 2. How do consumer reviews evolve over time?

We study the time-series property of online consumer reviews based on empirical data collected from Amazon.com to reveal why we suspect that vendors, publishers, and authors might consistently manipulate online reviews. Before we discuss our analytical and empirical models, we first discuss where and how we collected our data.

### 2.1. Data

We collected our data from Amazon Web Service (AWS) and constructed two datasets to examine our research questions. The first dataset is cross-sectional data composed of a random sample of books, DVDs, and videos. For this dataset, we collected the product information and corresponding consumer reviews from Amazon.com in July 2005.[3] The second dataset is a panel dataset composed of a sequence of online review information (price, sales, and review information) for a sample of books, DVDs, and videos collected over several months at approximately three-day intervals. The initial items in this panel dataset were randomly chosen from Amazon in July 2005. For the panel data collection, since it occurs approximately every three days, we identified each data collection batch by a unique sequence number. Because we need to know both the true product quality and the perceived product quality that consumers used to make purchase decisions, we used the

[1] http://www.washingtonpost.com/ac2/wp-dyn/A61073-2004Apr8?language=printer.

[2] According to http://www.clickfraudreport.com/1.html, the essence of click fraud is "any click where there is no intention by the clicker to purchase, browse or gain information from the website they visit. And the only goal of a click is to either to drain your marketing budget or generate revenue from the click". Even though we cannot find a dollar amount lose due to review fraud, we believe it is comparable to click fraud.

[3] This study is based on data collected in July 2005. We performed similar data analysis using data collected in February, March, and April of 2005, which rendered similar results.

**Table 1**
Summary statistics.

| Category | #Reviews | #Amazon items | #Distinct items | Average rating |
|---|---|---|---|---|
| *Panel A: Amazon cross-sectional data (July 2005)* | | | | |
| Book | 967,075 | 54,431 | 54,431 | 4.02 |
| DVD | 2,034,552 | 32,413 | 32,413 | 4.19 |
| Video | 1,248,992 | 44,489 | 44,489 | 3.99 |
| Total | 4,250,619 | 131,333 | 131,333 | 4.09 |
| *Panel B: Amazon panel data (July 2005–January 2006)* | | | | |
| Book | 6,759,764 | 261,187 | 10,052 | 3.87 |
| DVD | 4,056,340 | 258,736 | 9988 | 4.07 |
| Video | 4,371,833 | 259,736 | 10,000 | 4.02 |
| Total | 15,187,937 | 779,659 | 30,040 | 3.97 |

panel dataset to answer the questions as to whether consumers understand the existence of online review manipulation (Section 5) and how consumers make purchase decisions with the existence of review manipulation (Section 6). For the rest of the research questions, we use the cross-sectional datasets.

Because of some technical glitches in AWS, for the panel data, we had to exclude certain sequences in which only partial data were collected. For example, during several sessions, AWS did not respond to our queries or was offline and we were therefore able to process only partial or no data during these sessions. Each session is identified by a unique batch number. In total, we obtained 26 batches of review and item-level data. Table 1 provides summary statistics for our cross-sectional and panel data. On Amazon.com, consumers can report only an integer product review on a 1-star to 5-star scale, where 1-star = least satisfied and 5-star = most satisfied. The average review scores for books, DVDs, and videos are overwhelmingly favorable, reflected by the high average product reviews.

### 2.2. What does the order (relative time) mean?

To study how consumer reviews evolve over time, we first define a new term called "order" to represent the relative time. Order 1 means the first review every product received; Order 2 represents the second review every product received; and so on. In our study, we use relative time (order) instead of absolute time because each item sold on Amazon has its own release date and therefore its own absolute age on Amazon. This absolute age varies from 1 month to several years with a very large variance. Thus it is difficult to compare the change in review scores over time based on an absolute time. Since we are interested in the temporal properties of online reviews, using relative time allows us to pull items with different absolute ages together, under the assumption that reviews of different items have similar trends over time. In our later regression analysis, we did control the potential confounding effect of absolute time.

### 2.3. Potential drivers for reviews decrease over time

We first look at how consumer reviews change over time. Assume that there is no manipulation (online reviews are all given by previous customers) and no self-selection bias (later customers share the same tastes as early customers when evaluating the same product). Assume that a consumer will realize the true quality of a product after purchasing it and will truthfully report his/her opinion about that product. Thus, the online rating $r_{it}$ for product $i$ with $q_i$ at time $t$ is:

$$r_{it} = q_i + \varepsilon_{it} \tag{1}$$

where $\varepsilon_{it} \sim N(0, \sigma_{it}^2)$ represents the difference between the product's true quality and its online rating for product $i$ at time $t$. Thus, the average consumer rating for the $t$th review of all $K$ items is

$$\bar{r}_t = \frac{1}{k}\sum_{i=1}^{K} r_{it} = \bar{q} + \varepsilon_t, \text{ where } \bar{q} = \frac{1}{k}\sum_{i=1}^{K} q_i \text{ and } \varepsilon_t = \frac{1}{k}\sum_{i=1}^{K} \varepsilon_{it}. \tag{2}$$

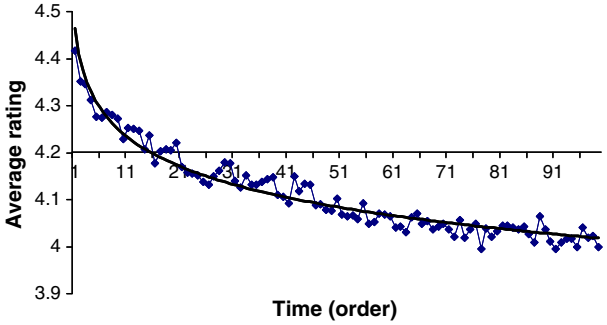


**Fig. 1.** Mean consumer ratings ($\bar{r}_t$) over time (order).

It is obvious that $\bar{r}_t$ should not change over time with the above assumptions. To test whether this is true, we chose those items out of our cross-sectional data that have received more than 100 consumer ratings[4] to make sure that these items have been in the market long enough to demonstrate their temporary pattern. We ended up with 1526 books, 2231 DVDs, and 2763 Videos. We estimate the average ratings of all items at each order (as Eq. (2)) and present the results in Fig. 1.

Fig. 1 shows that $\bar{r}_t$ decreases (with decreasing rate) with elapsed time. This raises the question: Why do the mean ratings decrease over time? Two potential drivers might be able to explain this kind of phenomenon:

- There are systematic differences between early customers and later customers, consistent with the higher-taste-self-selection theory proposed by Li and Hitt [9]. Normal consumer reviews of early periods are systematically positively biased, which leads to a decrease trend over time [9]. In addition, the researchers document that consumers are not fully rational because they do not fully correct for the review bias that occurs due to the self-selection. However, one embedded assumption in their paper is that there is no review manipulation and all reviews are truthful.
- There is a systematic positive manipulation from the vendors, publishers, and/or authors of the online reviews. This positive manipulation decreases with elapsed time as well, resulting in a decreasing trend of reviews over time. The reason for the positive manipulation bias at the early stage is linked to the cost and benefit of manipulation. Normally when an item is first available to an E-commerce market, there are very few consumer reviews, so the manipulation cost at this stage is relatively low because vendors need to write only a few reviews to change the mean consumer reviews. Also, vendors, authors, and publishers have higher incentives to engage in manipulating online reviews at this stage as well because it is at this phase that reviews have the highest impact on sales [8]. As time passes, this product will receive a large number of authentic consumer reviews. Under such a scenario, the cost to manipulate the outcomes of consumer reviews becomes very high.[5] From this point on, we assume that the likelihood of publishers, authors, and vendors manipulating online reviews decreases over time.

Therefore there are two competing processes that might cause mean consumer rating to decrease over time (Fig. 1). These two possibilities paint two different pictures. One believes that all reviews are truthful, while the other one hypothesizes that some reviews are

[4] Changing the cut-off point to another number, such as 130, 120, 110, etc., does not change our results qualitatively. We also make sure that the exact items included for each order are the same so that what we observed is really associated with the time instead of the item difference.

[5] Since a product receives a lot of reviews, vendors need to post a decent number of biased reviews to fight with unfavorable reviews and make the manipulation work. On a lot of websites, this can be very time consuming and costly.

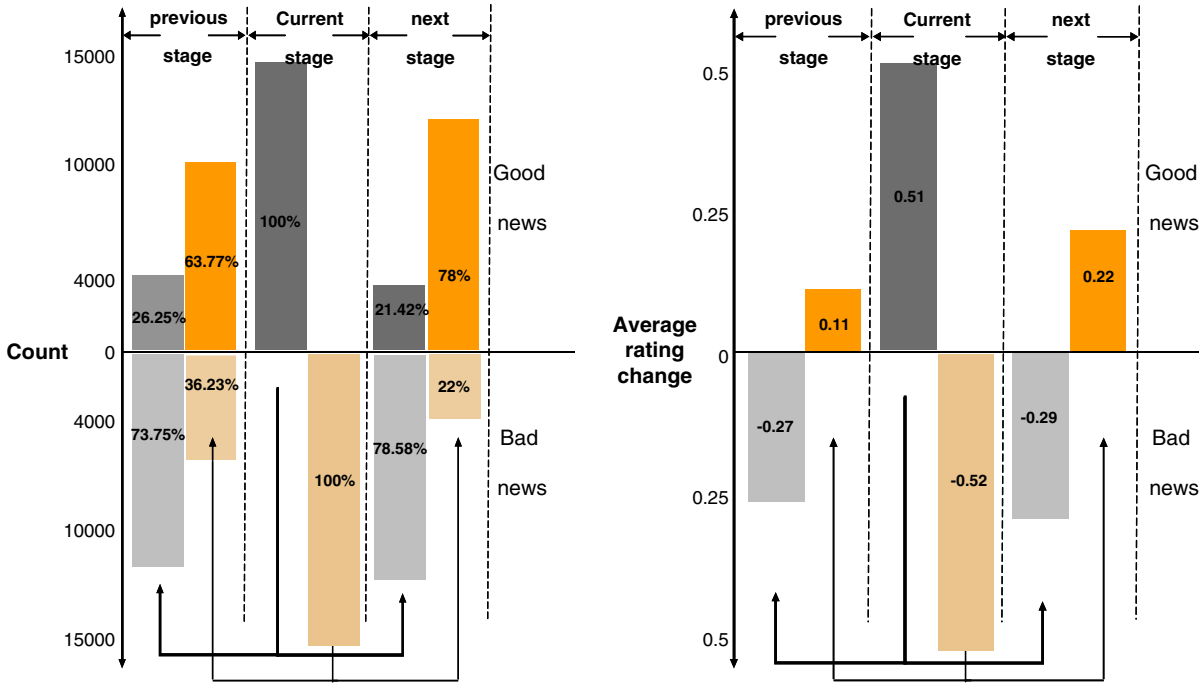


**Fig. 2.** Mean-reverse property of online reviews.

manipulated; one proposes that consumers are not able to fully correct self-selection bias, while the other one believes that consumers are smart enough to filter out the review manipulation. Is self-selection alone sufficient to explain the phenomena that the mean consumer rating decreases over time as Li and Hitt [9] observed? Are consumers really as smart as suggested by Dellarocas [2]?

Before we pursue the answer, let's first show why we suspect that there might be another potential driving force besides self-selection bias: manipulation with decreasing magnitude over time.

### 2.4. Why do we suspect that there might be review fraud?

To find out why we suspect that there might be systematic manipulation from book publishers, sellers, and/or authors in online product reviews, we adopt a portfolio approach using online review information from Amazon.com. The meaning of a portfolio in our context is different from that of a traditional finance context, where a portfolio represents a basket of securities typically designed to reduce risk. Our portfolio here is comprised of products and events (good and bad) that share similar characteristics.

There are two event types of interest in this study: good news events and bad news events. In our context, a good (bad) news event occurs when the newly released review for an item has a higher (lower) score than the average of its previous review scores. We are interested in knowing for those items that received good (bad) news from time $t-1$ to $t$ (*current period*), generally how will their product ratings be changed from time $t$ to $t+1$ (*future period*), and generally how were their product ratings changed from time $t-2$ to $t-1$ (*previous period*).[6]

Fig. 2 shows that out of all the items that received good reviews in the current period, 78.58% will receive bad reviews in the future period, while 73.57% received bad reviews in previous period. For all the items that received bad reviews in the current period, 78% will receive good reviews in the future period, while 63.77% received good reviews in the previous period; For those items included in the good news group, on average their mean consumer rating increases by 0.51 as we move forward from $t-1$ to $t$ (current period), on average their mean consumer rating decreased by 0.29 in the previous period and will decrease by 0.27 in the future period again. We also observe similar pattern for the items included in the bad news group. For such items, on average, mean consumer rating decreases by 0.52 from $t-1$ to $t$ (*current period*), however their mean consumer rating on average increased by 0.11 in the previous period (from $t-2$ to $t-1$), and will increase by 0.22 in the future period again (from $t$ to $t+1$).

In general, Fig. 2 reveals that online reviews demonstrate a mean-reverse property. That is given a decrease of consumer ratings of a product in the current period, there will, generally, be an increase in consumer ratings in the future period and vice versa. One possible explanation for such reverse property of online reviews is the existence of review manipulation. Online publishers, book authors, and vendors are continuously monitoring online reviews, and these entities will write strong positive reviews to boost the online consumer ratings whenever there is a decrease in consumer ratings.[7] However, future reviews written by new consumers will correct that manipulation and reverse the direction of the reviews.

In the Amazon market, there are other factors that might influence the changes in the reviews, such as the popularity of an item. To tease out the potential confounding influence of the popularity of a product, we classify our items into 10 equally spaced groups based on product sales ranks and repeat the same analysis. Our results show that even after controlling popularity, online consumer reviews still demonstrate a mean-reverse property.

Another potential reason for this mean-reverse property is the limitation in the review scores that consumers can give. Recall that at

[6] Again, for simplicity, we exclude those cases when review scores do not change. Our results stay the same regardless of whether we include or exclude no-change cases.

[7] Experienced products, such as books, might not be perfect substitutes for each other. Hence, we believe it is more feasible for a vendor to post strong positive reviews for his own products instead of posting negative reviews for his competitors.

Amazon, consumers can report only an integer product review score with a 1-star to 5-star Liker-type scale. In such a case, because later consumers cannot leave ratings of less than 1 or larger than 5, for items whose average rating is 1 (5) in the current period, their average ratings will definitely increase (decrease) in the future period and result in a mean-reverse phenomena even without publishers/authors/vendors' manipulation. We deleted those events where the mean of the consumer rating at the current period was 1 or 5 and repeated the same data analysis, for which we ended up with the same conclusion.

In addition, for the cross-sectional data, Fig. 3 shows that out of all the ratings that products receive (1, 2, 3, 4, and 5 in Amazon), only the percentage of "5" ratings decreases over time. The percentages of all of the other ratings increase in the very same manner and the relative magnitudes among the percentages of "1," "2," "3," and "4" ratings stay the same. One possible explanation of the large percentage of "5" ratings at the early stages of a product's release is manipulation. As time moves on, vendors are less likely to be involved in manipulation due to the increasing manipulation costs, thus the percentage of "5" ratings decreases over time.

Having said that, we understand that instead of being an indication of manipulation, another potential explanation for what we observed (Fig. 2) might be that online reviews follow a slow convergence process (self-selection) toward their associated true product quality. Hence, in our next section, we seek to uncover the real underlying driver.

## 3. Theoretical analyses: A pure self-selection process or a combination of self-selection with manipulation?

We now know that there are two potential drivers that might explain why reviews of most products tend to fall over time (Table 2). Our next question is whether a pure self-selection process without manipulation can be the sole underlying driving force. If a pure self-selection process is not the sole driver, then we are faced with a situation where self-selection and manipulation exist simultaneously. As we can see in Table 2, while manipulation might increase or decrease over time, so does self-selection. "High to low" (low to high) self-selection means that customers who have higher (lower) valuation of an item come in early (later), resulting in a positive (negative) but decreasing (increasing) bias over time. And the majority of consumer reviews of early periods are systematically positively biased [9]. As elaborated in Section 2, the likelihood of publishers, authors, or vendors coming in and manipulating the reviews decreases over time as well. Putting these together, Zone 1 in Table 2 is the most likely situation. Note that we do not assume that manipulation or self-selection of every item have a decreasing trend

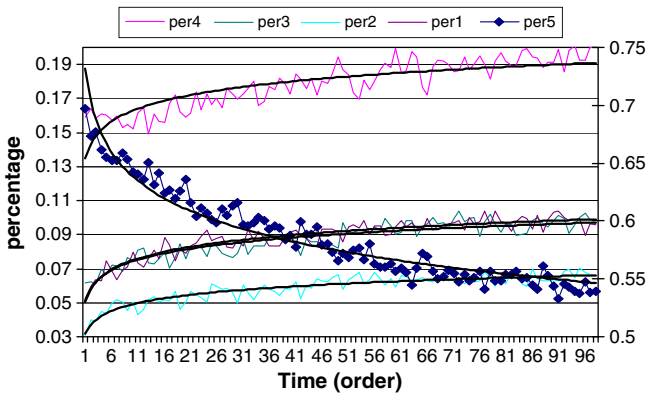


**Fig. 3.** Percentage of ratings over time.

**Table 2**
Driving force for online reviews over time.

| Self-selection | Manipulation | |
|---|---|---|
| | High → Low | Low → High |
| High → Low | Zone 1 | Zone 2 |
| Low → High | Zone 3 | Zone 4 |

over time. Our results still hold as long as the majority of the items follow a decreasing trend. From now on, we focus our research for Zone 1.

To test whether self-selection alone is sufficient to drive the temporal effect we observe, out of our cross-sectional sample, we select those books, DVDs, and videos that have at least 100 reviews. We then divide the reviews of each item into two subgroups. Group 1 includes all the reviews collected right after an item was released to the Amazon market (the first 25 reviews, excluding the first 5 reviews). The first 5 reviews are excluded because these reviews might be either randomly generated reviews or highly manipulated reviews. Please note that excluding the first 5 is a more conservative test of review manipulation, and even with these reviews included, qualitatively our results do not change. Group 2 includes, for the same group of items, the 81st review to the 100th review an item received. So group 1 represents the period in which higher manipulation or higher self-selection bias is more likely to occur. Group 2 represents the time period when the rating bias or manipulation bias is much less likely (near zero). Comparing the behavior of these two groups will enable us to identify the underlying drivers of the temporal effect.

### 3.1. The model: A pure self-selection process with "higher rating" consumers entering first

For this model, the underlying driving force is self-selection with "higher rating" consumers coming in first. "Higher rating" consumers refer to early adopters who are more enthusiastic about a product and who are more likely to leave positive reviews. In such a case, at time $t$, there will be a systematic self-selection positive bias $h_{it}$ incorporated within the online reviews $r_{it}$ with respect to the true product quality $q_i$ for product $i$ (Eq. (3)).

$$r_{it} = q_i + h_{it} + \varepsilon_{it} \tag{3}$$

For the majority of the products, we assume that $h_{it}$ is positive but decreases over time. $\varepsilon_{it} \sim N(0, \sigma_{it}^2)$ represents the difference between the online rating and a product's true quality. Thus, the average rating of $K$ products at the same time (order) $t$ is

$$\bar{r}_t = \frac{1}{K}\sum_{i=1}^{K} r_{it} = \frac{1}{K}\sum_{i=1}^{K}(q_i + h_{it} + \varepsilon_t) = \bar{q} + \bar{h}_t + \varepsilon_t \tag{4}$$

$\bar{q}$ is the average quality of all the $K$ products in our sample. Because the majority of $h_{it}$ is positive but decreasing over time (positive bias introduced by early adopters), thus $\bar{h}_t$ decreases over time (resulting in Fig. 1). As time goes on (when $t \to \infty$, $h_{it} \to 0$ and $\bar{h}_t \to 0$, thus $r_i \to q_i$ and $\bar{h} \to \bar{q}$), there will be no rating bias and online reviews will converge to products' true quality.

Define $\bar{R}_{i/t} = \frac{1}{t}\sum_{j=1}^{t} r_{ij}$ as the average consumer rating item $i$ received at the time period $t$, then

$$\bar{R}_i|_{t-1} = \frac{1}{t-1}\sum_{j=1}^{t-1}(q_i + h_{it} + \varepsilon_i) = q_i + \bar{h}_i|_{t-1} + \varepsilon_i'. \tag{5}$$

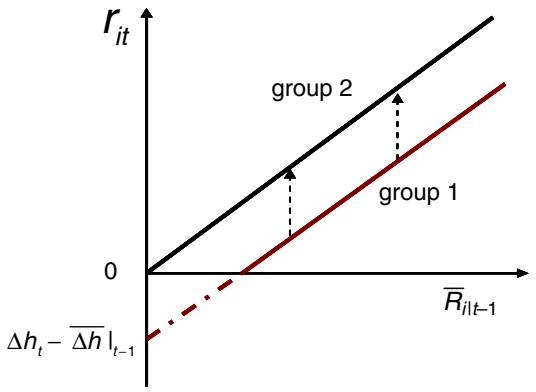


**Fig. 4.** Predicted relation between ratings and average rating based on positive rating bias assumption.

Taking Eq. (5) to Eq. (3), we can get the equation

$$r_{it} = \left(h_{it} - \overline{h}_{i|t-1}\right) + \overline{R}_{i|t-1} + \varepsilon''. \tag{6}$$

We assume that: 1) the higher-taste-self-selection bias decreases with a convex function over time; 2) self-selection is independent of quality; 3) quality is larger than the self-selection bias, namely *quality* >> $h_t - \overline{h}_{i|t-1}$). With the above assumptions, the difference between $h_t$ and $\overline{h}_{i|t-1}\left(h_t - \overline{h}_{i|t-1}\right)$ is smaller than zero for both the first group and the second group. However, that difference is bigger for the second group than for the first group. Furthermore, the slopes of both groups should be the same (Fig. 4).

**Proposition 1.** *If the underlying driver is self-selection bias (higher rating first), for the linear relation between the ratings of the current period and the average ratings of the previous period, group 1 and group 2 have the same slope (which equals 1: perfect positive linear correlation), but different intercepts. The intercept of group 2 is bigger than the intercept of group 1, but both are negative.*

### *3.2. The model: Pure manipulation (with decreasing likelihood of manipulation over time)*

The goal of manipulation behavior is to boost a product's online reviews in order to influence consumers' purchase decisions. We assume that a product will receive a review from an actual customer $1-\rho_t$ percent of time (assuming consumers know the true quality of the product of consumption and always truthfully report their evaluations when they write reviews). $\rho_t$ percent of time, that product will receive a manipulated review. Whenever vendors decide whether they should engage on online review manipulation, they need do a cost–benefit analysis. Since as time progresses products will receive an increasing number of authentic online consumer reviews, it becomes more difficult and costly to manipulate consumer opinions, thus we believe that $\rho_t$ decreases with elapsed time. $\overline{R}_{i|t-1|}$ is the average consumer rating item $i$ received at the time period $t-1$, while $\pi\left|A-\overline{R}_{i|t-1}\right|$ captures the incentive of the manipulation. $\pi$ is a standardized parameter, and $A$ reflects who is more likely to engage in manipulation.

Recall that on Amazon.com, 1-star = least satisfied and 5-star = most satisfied. Depending on the average rating a product received in the previous period, a vendor selling that product can decide whether to engage in manipulation at that time. For simplicity, we assume that for a firm deciding to adopt manipulation techniques, the actual manipulation strategy either monotonically increases or decreases with respect to the average rating a product received in the previous period, which is $\pi\left|A-\overline{R}_{i|t-1}\right|$. $A=5$ ($A=1$) represents the scenario where firms selling products with lower (higher) average consumer ratings are more likely to practice manipulation. We will find out which kind of manipulation happens on Amazon in the next section.

$$r_{it} = \begin{cases} q_i + \varepsilon_i^{\mathrm{c}} & (P_{\mathrm{c}} = 1-\rho_t) \\ \overline{R}_{i|t-1} + \pi\left|A-\overline{R}_{i|t-1}\right| + \varepsilon_i^{\mathrm{m}} & (P_{\mathrm{m}} = \rho_t) \end{cases} \tag{7}$$[8]

(Note $\varepsilon_i^{\mathrm{c}} \sim N(0, \sigma_{\mathrm{c}}^2), \varepsilon_i^{\mathrm{m}} \sim N(0, \sigma_{\mathrm{m}}^2)$, and c represents consumer and m represents manipulation).

So at any time (order) $t$, the expectation of $r_{it}$ is $E(r_{it}) = (1-\rho_t)q_i + \rho_t\overline{R}_{i|t-1} + \rho_t\pi\left|A-\overline{R}_{i|t-1}\right|$. Since the expectation and variance of $r_{it}$ are finite, for a group of items including $K$ number of products, based on the law of large numbers, the difference between the average ratings and expected ratings is finite

$$\frac{1}{K}\sum_{i=1}^{K} r_{it} - \frac{1}{K}\sum_{i=1}^{K} E(r_{it}) \sim N\left(0, \sigma_t^2\right). \tag{8}$$

Thus, the average ratings of these groups of items at the same time (order) $t$ is

$$\begin{aligned} \overline{r}_t &= \frac{1}{K}\sum_{i=1}^{K} E(r_{it}) + \varepsilon_t = \frac{1}{K}\sum_{i=1}^{K}\left[q_i + \rho_t\overline{R}_{i|t-1} - \rho_t q_i + \rho_t\pi\left|A-\overline{R}_{i|t-1}\right|\right] \\ &\quad + \varepsilon_t = \overline{q} + \frac{\rho_t}{K}\sum_{i=1}^{K}\left(\overline{R}_{i|t-1} - q_i\right) + \frac{\rho_t\pi}{K}\sum_{i=1}^{K}\left|A-\overline{R}_{i|t-1}\right| \\ &\quad + \varepsilon_t \overline{r}_t = \overline{q} + \omega\rho_t + \varepsilon_t \end{aligned} \tag{9}$$

where,

$$\omega = \frac{\rho_t}{K}\sum_{i=1}^{K}\left(\overline{R}_{i|t-1} - q_i\right) + \frac{\rho_t\pi}{K}\sum_{i=1}^{K}\left|A-\overline{R}_{i|t-1}\right| > 0 \tag{10}$$

$\omega$ is a finite positive number. And as $t \to \infty, \rho_t \to 0$, thus, as time moves on, the average rating of the $t$th reviews over all $K$ items is also decreasing over time, resulting in Fig. 1 as well

- If the products with higher average ratings are more likely to be manipulated ($A$ is 1), then.

$$\begin{aligned} E(r_{it}) &= (1-\rho_t)q_i + \rho_t\overline{R}_{i|t-1} + \rho_t\pi(\overline{R}_{i|t-1} - 1) \\ &= -\rho_t\pi + (1-\rho_t)(q_i - \overline{R}_{i|t-1}) + (1 + \rho_t\pi)\overline{R}_{i|t-1} \end{aligned}$$

  $q_i - \overline{R}_{i\,t-1}$ should be related to $\rho_t$; the larger the $\rho_t$, the bigger the difference between its quality and its average rating. Thus, $E(r_{it}) = -\rho_t\pi - \theta\rho_t(1-\rho_t) + (1 + \rho_t\pi)\overline{R}_{i|t-1} = \lambda_1\rho_t + (1 + \rho_t\pi)\overline{R}i|_{t-1}$ where $\theta > 0$ and $\lambda_1 = -\rho_t\pi - \theta\rho_t(1-\rho_t) < 0$. So for the first group, because $\rho_t > 0, \lambda_1\rho_t < 0$ (intercept) and $1 + \rho_t\pi > 1$ (slope); while for the second group, because $\rho_t \to 0, \lambda_1\rho_t \to 0$ (intercept) and $1 + \rho_t\pi \to 1$ (slope).
- If the products with lower average ratings are more likely to be manipulated ($A$ is 5), proceeding along lines similar to the above, we obtain the following expression: $E(r_{it} = \lambda_2\rho_t + (1-\rho_t\pi)\overline{R}_{i|t-1}$. Under such a circumstance, for the first group, $\lambda_2\rho_t > 0$ (intercept) and $1 - \rho_t\pi < 1$ (slope); while for the second group, because $\rho_t \to 0, \lambda_2\rho_t \to 0$ (intercept) and $1 - \rho_t\pi \to 1$ (slope).

Combining the above cases, one can draw the conclusion that when the underlying driver is manipulation, the plot of the current

[8] This is a strategy that is more feasible for vendors to implement because for experienced goods, quality is non-verifiable before the goods are consumed. This means vendors should manipulate based on the market signal (reviews) both they and consumers can observe. Later, we investigate the ways in which vendors consider both quality and reviews to make manipulation decisions.

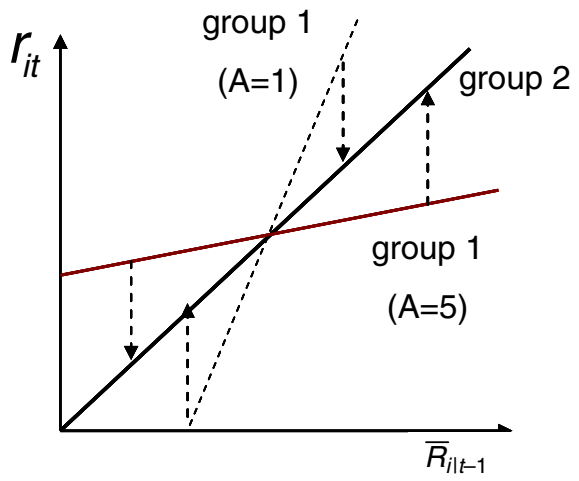


**Fig. 5.** Predicted relation between ratings and average rating based on manipulation assumption.

rating against its previous average rating for groups 1 and 2 is different for both slope and intercept (demonstrated in Fig. 5).

**Proposition 2.** *If the underlying driver is manipulation, assuming that manipulation decreases over time, for the linear relation between the ratings of current period and the average ratings of previous period, group 1 and group 2 have different slopes (the slope of group 2 is close to 1) and different intercepts (the intercept of group 2 is close to 0). In addition, if products with higher (lower) average ratings are more likely to be manipulated, then the slope of group 1 is bigger (smaller) than 1 and the intercept of group 1 is smaller (bigger) than 0.*

### 3.3. If there is no "positive rating bias driven by self-selection bias" and no "manipulation"

When there is no "self-selection" or "manipulation" involved, we can obtain:

$$\overline{R}_{i|t-1} = \frac{1}{t-1}\sum_{j=1}^{t-1} r_{ij} = \frac{1}{t-1}\sum_{j=1}^{t-1}(q_i + \varepsilon_i) = q_i + \varepsilon_i'. \tag{11}$$

Further, taking Eq. (11) and substituting it into Eq. (1), we have

$$r_{it} = \overline{R}_{i|t-1} + \varepsilon''. \tag{12}$$

Under such an assumption, both the first group and the second group have the same intercept and slope.

### 3.4. The empirical test and robustness check

We regressed the rating on the lag average rating for reviews of group 1 (orders 6–25) and group 2 (orders 81–100) separately. Due to the nature of this sample, we expect that manipulation is more likely for the first group. Because we include reviews of various products over time, we must control the heterogeneity of age, popularity, or reviewer characteristics over time and across different product items. However, we cannot control such heterogeneity by running a fixed effect model at individual item level because doing so for group 1 and group 2 will definitely result in different slope and intercept estimations for these two groups. Auto regression with trend will perturb our result. Thus, in order to get the right estimation, we control the following potential confounding factors:

- Age and time effect
  Each group includes products with age differences (age refers to how long a product has been released to the Amazon market), and for each product, it includes reviews belonging to 20 orders. Thus, the self-selection pattern might vary over different products with different time. To control for the age and time effect, we add two variables, namely *Lag*(*log*(*T*)) and *DifT*. *Lag*(*log*(*T*)) is used to control the age of the review at the previous period. For a particular review written for one specific item, it is estimated as the date difference between the previous review date and the date that item was released to the Amazon market. *DifT* is used to control the number of days difference between the current review and its nearest previous review. To summarize, *Lag*(*log*(*T*)) is used to control the self-selection time characteristics at time $T-1$, while *DifT* is used to control the self-selection time characteristics of the current rating.
- Silence
  For each item, we construct one variable termed "*Silence*" (*Silence* = *T*/# of Reviews) to control for product popular effect. *Silence* represents the mean inter-arrival time between two adjacent reviews. A smaller *Silence* value represents an item with higher popularity.
- Reviewer quality change over time
  The role of reviewers also influences how consumers act upon online reviews. If one product received a higher percentage of expert reviews, then generally its reviews will be more useful with less "self-selection" (Because the experts understand more about the quality of the product, their reviews will have less bias). How to distinguish which reviews are more professional? For every review posted on Amazon.com, it provides the data about how many other customers read that review (*totalvotes*) and how many think that review is helpful (*helpfulvotes*). Thus we define a variable called *Helpration* (*Helpration* = # of *helpfulvotes*/# of *totalvotes*). For each item and at time *T*, we estimate the mean of the *Helpful* ratio of all reviews received at time $T-1$ for that item, termed *Lag*(*AvgHelpratio*), to control the change of review quality over time. Lastly, we add the *DVDdummy* and the *Vhsdummy* to control the product category. The final model is as follows

$$\begin{aligned} Rating = {} & \beta_0 + \beta_1 Lag(Avgrating) + \beta_2 Lag(Log(T)) + \beta_3 Lagdif T \\ & + \beta_4 Lag(Popularity) + \beta_5 Lag(Avghelpratio) \\ & + \beta_6 Dvddummy + \beta_7 Vhsdummy + \varepsilon \end{aligned} \qquad \text{Model 1}$$

Based on what the real underlying driver is, manipulation or self-selection, we expect to see the following results in Table 3. By testing whether $\beta_1$ of group 1 equals to that of group 2, we can uncover which is the real driver, self-selection or manipulation.

Furthermore, we conduct a White test to check the existence of heteroscedasticity and cannot accept the homoscedasticity at the 5% level. Based on the procedure proposed in Long and Ervin [10], we run a SAS macro to correct the potential heteroscedasticity problem. Qualitatively the results didn't change.

Table 4 presents the regression results for these two groups. The intercept of group 1 (Para = 1.16, p-value < 0.0001) is much bigger than that of group 2 (Para = 0.14, p-value < 0.63). The slope of group 1 (Para = 0.71, p-value < 0.0001) is also different from that of group 2 (Para = 0.96, p-value < 0.0001). In addition, the slope of group 1 (0.71) is significantly smaller than 1, while the slope of group 2 (0.96) is not significantly different from 1. If the underlying process for the majority of the items is self-selection (AR with trend), then we

**Table 3**
Expected regression results with different divers.

| Drivers | | Group 1 | Group 2 |
|---|---|---|---|
| Higher average rating → higher manipulation | Intercept | <0 | =0 |
| | Slope | >1 | =1 |
| Lower average rating → higher manipulation | Intercept | >0 | =0 |
| | Slope | <1 | =1 |
| Self-selection bias (positive rating bias) | Intercept | <0 | <0 |
| | Slope | 1 | 1 |

**Table 4**
The relation between ratings and lag average rating.

| Variable | Group 1 | Group 2 |
|---|---|---|
| *Intercept* | 1.16*** | 0.14 |
| | (38.69) | (0.63) |
| *Lagavgrating* | 0.71*** | 0.96*** |
| | (139.62) | (157.84) |
| *Laglogt* | −0.03*** | −0.02 |
| | (−3.78) | (−0.47) |
| *Logdift* | 0.02*** | 0.01*** |
| | (6.72) | (4.86) |
| *Logpopularity* | 0.04*** | 0.03 |
| | (5.34) | (0.56) |
| *Lagavghelpratio* | −15.11* | −1.60 |
| | (−1.66) | (−0.91) |
| *Dvddummy* | 0.01 | 0.01 |
| | (1.25) | (0.87) |
| *Vhsdummy* | −0.02*** | 0.01 |
| | (−2.63) | (1.01) |
| N | 130,400 | 130,400 |
| Adj R square | 0.15 | 0.17 |

*** $P \leq 0.01$, * $P \leq 0.1$.

should expect the slope of group 1 to be exactly the same as that of group 2 (see Table 3). However, this is not what we observe in Table 4. Combining the above results, we believe that a pure "Self-Selection bias" cannot lead to the phenomena we observe. We conclude that these results reveal the existence of positive manipulation behavior, and prove that products with a lower average rating are more likely to be manipulated. At the same time, it is worth noting that the "Self-Selection bias" proposed by [9] cannot totally be ruled out. At the very least, we prove that manipulation must be present in order to drive such phenomena. This might indicate that what we observed is the result of joint forces: manipulation and self-selection. For robustness check, we also consider 3 other cases, such as running regression using cross-sectional data. Overall we still observe the existence of manipulation. Please refer to Appendix A for details.

## 4. Relation between quality and manipulation[9]

In Section 3, we studied the relation between average rating and manipulation, and documented that reviews of products with low average ratings would be more likely to be manipulated. In this section, we investigate the relation between product quality and manipulation because the average rating of a product is not necessarily the same as its product quality, especially for the early stage reviews. Therefore we seek to answer the following question: what kind of products is more likely to be manipulated, low-quality products (with lower average ratings) or high-quality products (with lower average ratings)? Or do low- and high-quality products share an equal chance to be manipulated? Dellarocas [2] pointed out that online reviews are more informative if every firm's manipulation strategy is a monotonically increasing function with respect to that vendor's true quality. However is this the strategy that every manipulator really adopted?

In order to find the answer to the above question, we study the variance of the quality w.r.t. average consumer rating at orders 7, 27 and 87, where we expect to see high, low, and nearly no manipulation occurring in these three periods respectively. Also, as the time (order) elapses, the uncertainty of the consumer reviews will also go down and converge to the products' true quality.

If there is no manipulation (Fig. 6A), then, the variance of the quality with respect to the average rating should be very close to 0.

If every vendor has an equal chance to engage in manipulation and the vendors' manipulation strategy is fully systematic (see Fig. 6B), the variance of the quality with respect to the average rating should remain constant. Meanwhile, for any given average rating, with the elapsed time, the variance of quality with respect to that average rating will decline (The light gray area will become narrow).

However, some firms might decide not to be involved in manipulation from day 1 because they care more about their own reputations, or they have limited resources that prevent them from engaging in such activities, or the product in question is not one of the mainstream products of that vendor. Thus, at the same time (order), even for different vendors selling the same quality of products, the probability of manipulation $\rho_t$ may be different. We assume that $\rho_{it}$ is uniformly distributed between 0 and max $\rho_{it}$. If different vendors selling products of the same level of quality indeed adopt different manipulation strategies, quantified by different $\rho_{it}$, then for products receiving the same average rating at the same order $t$, their quality will be different. The variance of the quality w.r.t average rating will be significantly greater than 0 (see Fig. 6C and D). The dark gray area represents the manipulation zone, while the light gray area represents the noise. Moreover, as time (order) goes by, $\rho_t \to 0$, no matter whether the manipulation strategies are the same among different vendors, there will be no manipulation at the end. The variances of quality w.r.t to average ratings should be converged to 0 (see Fig. 6A3, B3, C3, and D3).

Given that the existence of manipulation has been proven in Section 3, within the same period (order 7 and order 27),[10] we develop the following hypotheses:

**H1a.** If reviews of lower quality products (with lower average ratings) are more likely to be manipulated, within the same period when the average rating increases, the variances of quality will at least demonstrate a decreasing trend.[11] In addition, across different periods, that variance will converge to zero with elapsed time (see Fig. 6D1–2).

**H1b.** If reviews of higher quality products (with lower average ratings) are more likely to be manipulated, within the same period when the average rating increases, the variances of quality should not decrease. Furthermore, across different periods that variance will converge to zero with elapsed time (see Fig. 6E1–2).

To test our Hypotheses 1a and 1b, out of our cross-sectional sample, we still select those books, DVDs, and videos that have at least 100 reviews.[12] For these items, the average consumer rating at order 100 is chosen to be a measurement of a product's true quality.[13] We also try other ways to define quality, such as the method used in Section 5. Qualitatively the results do not change. For each item, we calculate its average rating at time (order) $t$ followed by an estimation of the variances of the quality. In order to avoid the potential issue caused by the rating bound,[14] we focus on studying only the average rating between 1.5 and 4.5. Fig. 7 shows that as the average rating goes up, the variances of quality go down (order 7 and order 27). As time elapses, the

[9] Here "higher quality vendors" actually means vendors selling products with a higher quality.

[10] At order 87, there is almost no manipulation.

[11] Technically speaking, we should see an increasing trend followed by a decreasing trend. However, depending on the manipulation level and the data, the increasing trend part might not be observable.

[12] The number of 100 reviews was chosen to ensure the items have been in Amazon long enough to go through the high manipulation (self-selection) stage to low manipulation (self-selection) stage. At this time (order 100), the average rating is a good proxy for the product's true quality.

[13] The slopes of group 2 for Table 4 and Table 5 are not significantly different from 1, indicating that at this time, the average rating already converges to the product's true quality. Furthermore, changing this to order 90, 110, etc. qualitatively does not change our final results.

[14] Because of this boundary, we might not be able to observe the variance goes up portions.

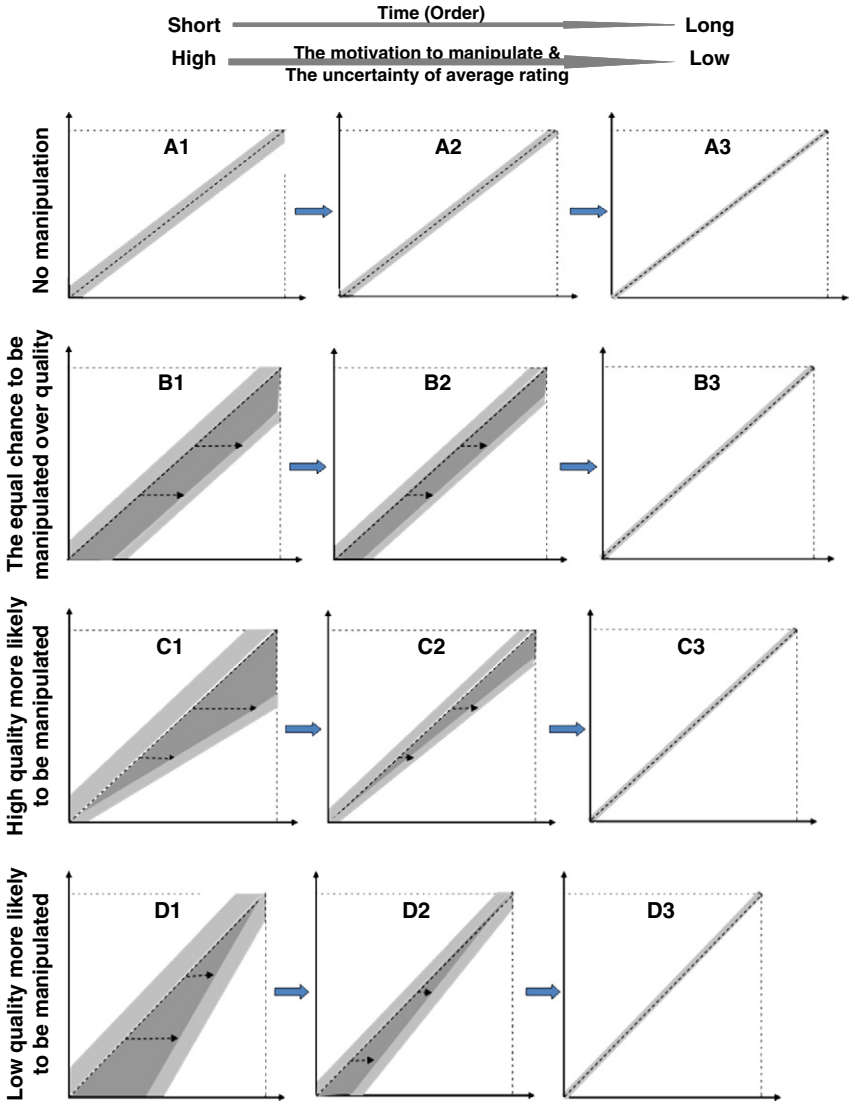


**Fig. 6.** Relationship between quality and average rating.

variances of quality decrease as well. And, at order 87 that variance is almost 0 (supporting our Hypothesis 1a). In general, our empirical results show that even vendors that sell products of the same quality adopt different manipulation strategies, and it is more likely for a vendor selling the lower quality products and receiving low average consumer ratings to engage in manipulation. This type of manipulation indeed makes things even worse because under such a circumstance, online reviews are much less informative.

## 5. Are consumers able to fully account for bias?

Our previous analyses suggest that there is systematic manipulation of online consumer opinions, but not every vendor engages in manipulation. Even vendors that sell products of the same quality adopt different manipulation strategies. As the manipulation strategy is not fully systematic, the higher quality products may show lower average ratings. In contrast, the lower quality products may exhibit higher average ratings (Fig. 8). Hence, there is a disconnection between quality and average rating. Under such a circumstance, we hypothesize that consumers might not be able to fully correct for this bias when making purchase decisions because they cannot tell which vendors are or are not manipulating online reviews. The best they can do to correct for this bias is based on an expected overall market manipulation.

We use a panel dataset instead of a cross-sectional dataset to study whether consumers fully account for the self-selection bias and manipulation. The reason for using a panel dataset is that in order to test this hypothesis, for each item we need to know its sales, review, and price information for the period when this item has high chance of being manipulated. In addition, for the same item, we also need to know such information for the period when there is almost no manipulation, such as its true quality.

Before we present our hypotheses, let's first introduce three key constructs. In order to test whether consumers fully correct the bias, out of the panel data sample, for each batch of data, we select those items whose total numbers of reviews at that batch level is less than 25. The number 25 was selected to ensure that serious manipulation or self-selection was more likely to be occurring and that the average ratings at that time did not reflect a product's true quality. We then collected the consumer reviews for these products again in January 2008. Those items whose numbers of reviews in January 2008 were still fewer than 65 were deleted from our sample to make sure that the items had received enough reviews and that their average ratings

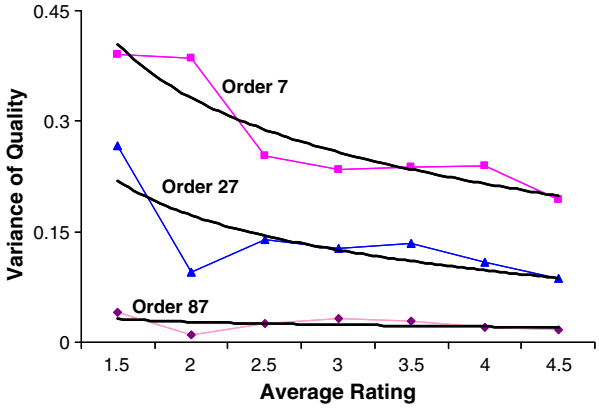


**Fig. 7.** Quality variance with respect to average rating.

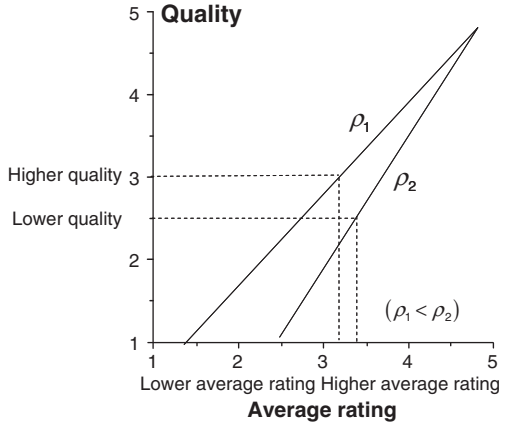


**Fig. 8.** Misalignment between average rating and quality.

at this point (2008) represented the true product quality.[15] The average rating collected in 2005 is called $\bar{r}^{ori}$; the one collected at 2008 is named $\bar{r}^{q}$ (approximated as the true product quality).

Because different vendors adopt different manipulation strategies, we expect that consumers might not be able to fully correct for the bias caused by vendors' manipulation strategies. The best consumers can do is to estimate the quality based on the expected overall manipulation level. So we can derive the rating consumers used to make a purchase decision (termed $\bar{r}^{adj}$) based on Model 2:

$$AvgRating - Quality = \beta_0 + \beta_1(5 - Quality) + \beta_2 Lag\log T + Lagdif T + \beta_4 LagAvghelpratio + \beta_5 Dvddummy + \beta_6 Vhsdummy + \varepsilon. \quad \text{Model 2}$$

The difference between the actual quality (the future average rating collected in 2008) and the historical average rating (collected in 2005) represents the quality bias either due to manipulation or self-selection. So, now the original average rating a product received can be broken into three components:

$$\bar{r}_i^{orig} = \bar{r}_i^{adj} + \left(\bar{r}_i^{q} - \bar{r}_i^{adj}\right) + \left(\bar{r}_i^{ori} - \bar{r}_i^{q}\right)$$

- $\bar{r}_i^{adj}$ represents consumers' estimated product quality according to their expectation of the overall manipulation and self-selection. Where

$$\bar{r}_i^{adj} = \frac{1}{1-\hat{\beta}_1}(AvgRating - \hat{\beta}_0 - 5\hat{\beta}_1 - \hat{\beta}_2 Lag\log T - \hat{\beta}_3 Lagdif T - \hat{\beta}_4 LagAvghelpratio - \hat{\beta}_5 Dvddummy - \hat{\beta}_6 Vhsdummy$$

- $\bar{r}_i^{ori} - \bar{r}_i^{q}$ measures the bias introduced by the manipulation and self-selection effects included in the original average rating.

- $\bar{r}_i^{q} - \bar{r}_i^{adj}$ has two possible interpretations. When $\bar{r}_i^{q} - \bar{r}_i^{adj}$ is greater (less) than zero, it represents the situation for a given item $i$: either consumers over-adjust (under-adjust) or vendors are more (less) honest and the manipulation level of that vendor is relatively smaller (bigger) than the overall market level manipulation.

**H2a.** If customers are able to fully correct for the bias, the sales of a product should be positively correlated with its true quality, approximated by its future average rating ($\bar{r}_i^{q}$). And the sales of a product should not be correlated with $\bar{r}_i^{ori} - \bar{r}_i^{q}$.

**H2b.** If customers can only partially correct such a bias, the sales of a product should be positively correlated with $\bar{r}_i^{ori}$, $\bar{r}_i^{adj}$ and $\bar{r}_i^{ori} - \bar{r}_i^{q}$. And the sales of a product should not be correlated with $\bar{r}_i^{q} - \bar{r}_i^{adj}$.

**H2c.** If customers can only partially correct such a bias, the sales of a product should be positively correlated with $\bar{r}_i^{ori}$, $\bar{r}_i^{adj}$ and $\bar{r}_i^{ori} - \bar{r}_i^{q}$.

We use the following three empirical models to validate our hypotheses and present the results in Table 5.

$$In(SalesRank_{i+1}) = \beta_{11}\bar{r}_i^{q} + \beta_{21}\log(SalesRank_i) + \beta_{31}\log(price_i) + \beta_{41}\log(Num\_rev_i) + \beta_{51} DVD\_Dummy + \beta_{61} VHS\_Dummy + \varepsilon_{i1} \quad \text{Model 3a}$$

$$In(SalesRank_{i+1}) = \beta_{12}\bar{r}_i^{ori} + \beta_{22}\log(SalesRank_i) + \beta_{32}\log(price_i) + \beta_{42}\log(Num\_rev_i) + \beta_{52} DVD\_Dummy + \beta_{62} VHS\_Dummy + \varepsilon_{i2} \quad \text{Model 3b}$$

$$In(SalesRank_{i+1}) = \beta_{13}\bar{r}_i^{adj} + \beta_{23}\left(\bar{r}_i^{q} - \bar{r}_i^{orig}\right) + \beta_{33}\log(SalesRank_i) + \beta_{43}\log(price_i) + \beta_{53}\log(Num\_rev_i) + \beta_{63} DVD\_Dummy + \beta_{73} VHS\_Dummy + \varepsilon_{i3}. \quad \text{Model 3c}$$

Recall that *SalesRank* is the opposite of sales. Model 3a of Table 5 shows that the future average rating (proxy for quality) is insignificantly (Para = −0.06, p-value = 0.1125) negatively associated with the historical sales rank, while $\bar{r}^{ori}$ (Model 3b) is significant (Para = −0.13, p-value < 0.0001). This indicates that consumers are not able to fully account for the bias caused by self-selection and manipulation. Model 3c of Table 5 shows that both the consumer adjusted quality $\bar{r}_i^{adj}$ (Para = −0.20, p-value < 0.05) and the manipulation and self-selection bias $\bar{r}_i^{ori} - \bar{r}_i^{q}$ (Para = −0.68, p-value < 0.005) are significantly negatively associated with the historical sales rank, while the

[15] To cross-validate whether the method we proposed is correct, we compare the predictive power of average ratings to sales by linking the historical average rating (collected in 2005) and future average rating (collected in 2008) to the sales in 2005 for those items that received at least 65 reviews (in 2005) respectively. The number 65 was selected to test whether the average rating of items with 65+ reviews is a good indication of a product's true quality. The historical rating and future rating demonstrate the same magnitude and significance in terms of predicting historical sales of 2005. Furthermore, when we run our regression (Model 1) using order 65 and 66, the coefficient of the *LagAvgrating* is 0.95 and is not significantly different from 1. The other control variables are not significant. So basically there is no "Manipulation" and "Self-selection" after order 65. In addition, we also tried other cutting points, such as 75, 85, 90, and 100, and got very similar results.

**Table 5**
Regression analysis result of whether consumers are able to fully account for manipulation bias (dependent variable: ln(*SalesRank*)).

| Parameter | Model 3a | Model 3b | Model 3c |
|---|---|---|---|
| *Intercept* | 0.78*** | 1.22*** | 1.67*** |
| *Quality* | −0.06 | | |
| *Original rating* | | −0.13*** | |
| *Adjusted rating* | | | 0.20*** |
| *Quality*−adjusted rating | | | −0.58 |
| *Original rating*−quality | | | −0.68** |
| *Lag log* (sales rank) | 0.92*** | 0.92*** | 0.92*** |
| *Lag log* (price) | −0.001 | −0.01 | −0.02 |
| *Lag log* (no. reviews) | 0.02 | 0.004 | −0.01 |
| *DVDdummy* | −0.06 | −0.06 | −0.06 |
| *Vhsdummy* | 0.01 | −0.03 | 0.02 |
| N | 1245 | 1245 | 1245 |
| R square | 0.86 | 0.87 | 0.87 |

*** P≤0.01, ** P≤0.05.

**Table 6**
Regression analysis result of the "price quality" indicator (dependent variable: ln (*SalesRank*)).

| Parameter | Sample group | | | |
|---|---|---|---|---|
| | >100 | [55,65] | [15,25] | [5,15] |
| Manipulation probability | Low----------->High | | | |
| *Intercept* | 0.71*** | 1.65*** | 0.94*** | 1.43*** |
| *Average rating* | −0.12*** | −0.1*** | −0.03* | −0.1*** |
| *Lag log* (sales rank) | 0.87*** | 0.9*** | 0.9*** | 0.9*** |
| *Lag log* (price) | 0.06** | 0.01 | −0.01 | −0.06** |
| *Lag log* (no. reviews) | 0.13 | −0.05 | 0.05 | −0.01 |
| *DVDdummy* | −0.02 | −0.18 | −0.10 | −0.09 |
| *Vhsdummy* | 0.08 | −0.04 | −0.02 | 0.03 |
| N | 2339 | 7810 | 3105 | 1805 |
| 0.81 | 0.81 | 0.86 | 0.84 | 0.84 |

*** P≤0.01, ** P≤0.05, * P≤0.1.

coefficient before the quality minus adjusted rating variable is not significant (p-value>0.10), indicating that even though consumers are able to adjust for the manipulation bias and self-selection bias, they can adjust for it only partially. Vendors are able to cheat consumers by manipulating the final outcomes.

## 6. How do customers make purchase decisions when manipulation exists?

In the above sections, we show that, to some degree, online reviews are not trustworthy. Under such a circumstance, what information do consumers use to make purchase decisions? For experienced goods sold through the online electronic marketplace, in the absence of review manipulation, consumer reviews can be considered as a superior quality signal because these online reviews providing information about an item's value are written by previous customers after consumption. However, at the early stage with the presence of review manipulation, the story is different. Reviews of this stage are no longer fully trustworthy and might be downgraded to an "inferior" quality proxy. When not every vendor manipulates online reviews and consumers cannot discern who is and who is not manipulating, a higher price might lead to an increasing instead of a decreasing demand because a higher price might emerge as a quality index.[16] This is consistent with previous literature that vendors can use price or advertisement to signal their products' quality.

In order to test the existence of the "price quality" indicator in Amazon, we run separate regressions using four different sub-samples out of our panel data. For each sub-sample and each batch of data, we select those items whose total numbers of reviews at that batch level are greater than 100, between 55 and 65, between 25 and 15, and between 5 and 15 respectively. As stated before, as time progresses, the manipulation will decrease. So we expect that the probability of manipulation will be larger for the sub-sample composed of items with early stage reviews than for the sub-sample composed of items with later stage reviews. At the same time, prices will change from being positively associated with sales (at the early stage, reviews are manipulated so that price becomes a better quality signal) to being negatively associated with sales. At that later stage, the manipulation will be almost zero. Price becomes dis-utility because for two products with the same quality, signaled by the same average rating, consumers will select the one with the lower price because to consumer that product has a higher net utility.

**H3.** The price and the product sales will move in the same way when the manipulation is present; in the absence of manipulation, an increase in price will lead to a decrease in sales.

[16] Here we assume that vendors know the true quality of a product and will charge the optimal price to maximize their profits.

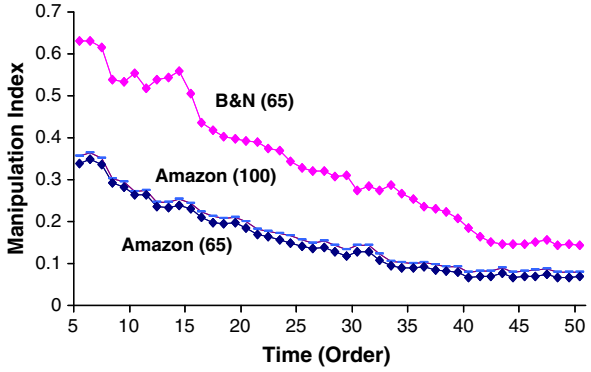


**Fig. 9.** Manipulation at different times for Barnes & Noble vs. Amazon.[17]

Table 6 shows that at the beginning, price is significantly (Para = −0.06, p-value<0.05) negatively associated with the future sales rank (proxy for the inverse of sales). This indicates that consumers use price as a quality signal because online reviews are less trustworthy when manipulation is present. The higher the price, the larger the sales will be (the lower the sales rank). As time progresses, the relationship changed to be negative and statistically insignificant, positive and statistically insignificant and finally positive and statistically significant (Para = 0.06, p-value<0.05). This supports our Hypothesis 3.

## 7. Manipulation across websites

Readers might think that our results are driven by special features of the data from Amazon. To check the robustness of our results, we collected 190,135 reviews for about 5149 items from Barnes & Noble in January 2008 and repeated the same analysis as in Section 3. Qualitatively, we arrived at similar results and found consistent manipulation of product reviews on Barnes & Noble as well.

Further, we estimated the overall manipulation levels at Barnes Noble and Amazon over time and plotted the results in Fig. 9. For each order (time), we run the regression based on Model 2 to get an estimation of $\beta_1$, which is used to approximate the manipulation index, representing the overall manipulation level. Our estimates show that on average, the manipulation levels decrease over time on both Barnes & Noble and Amazon. However, Barnes & Noble shows a higher level of manipulation. Our interpretation for such an observation is that Amazon has a greater wealth of consumer reviews and better reviewer qualities.

[17] For Amazon, we select those items with more than 100 reviews from our cross-sectional data. The average rating we retrieve at that time represents the true quality of those items. On Barnes & Noble, due to data limitations, the cut-off point is 65. We also use 65 reviews as a cut-off point for Amazon and arrive at similar results.

For example, Amazon has formed online clubs called "Purchase Circles" for people with similar interests in which reviewers and customers can build connections with each other through chat, discussion, and debate. Our results are not driven by the sampling issue because only 536 books fit our sample selection criteria (see Fig. 2 legend) when we estimate the manipulation level for Barnes & Noble, while 1526 books fit the criteria for Amazon. To make sure that the results are not driven by sample selection bias, we randomly selected 536 items out of the 1526 Amazon books and repeated the same analysis. Quantitatively and qualitatively, the results do not change.

Please note that the decreasing trend captured by Fig. 9 might be caused by the joint forces of manipulation and self-selection bias. It is reasonable to assume that the self-selection bias on Amazon is similar to that on Barnes & Noble because it is very unlikely that these two websites serve two groups of customers with completely different tastes over time. Thus, even with self-selection bias, we can still draw the conclusion that manipulation is a more serious problem on Barnes & Noble than on Amazon.

## 8. Discussions, conclusions, and future researches

In this study, we use data from Amazon and Barnes & Noble to document that publishers, authors, and vendors consistently manipulate online consumer reviews. If a firm decides to adopt manipulation, its manipulation strategy is monotonically decreasing with respect to that product's true quality. Under such a case, manipulation actually decreases the informativeness of online reviews. However, we prove that not all firms will manipulate online reviews. Because of this non-systematic involvement, it is not easy for consumers to fully correct for manipulation bias. Consumers can adjust for that bias based only on their expectations about overall manipulation rates. To some degree, vendors are able to manipulate the outcomes of the results and consumers therefore respond to the wrong information. We document the existence of the "price quality proxy" in the sense that at the early stage after an item is released to the Amazon market, consumers use price as a quality indicator instead of using the average rating. Thus, a higher price leads to an increase rather than a decrease in sales. Finally, we show that generally there is a higher level of manipulation on Barnes & Noble than on Amazon.

We document that the lower the quality and average rating of the products a vendor is selling, the higher the likelihood that that vendor is going to conduct online manipulation. This makes online reviews much less informative than when either there is no manipulation or when vendors selling higher quality products are more likely to manipulate online consumer opinions. This might result in consumers' totally discarding online reviews, defying the purpose of vendors' building online review systems and providing customers with an online review option. Over the long run, online markets such as Amazon.com or Barnes & Noble.com cannot maintain the quality of their online consumer opinion information when such manipulation is taking place. If the market continues to evolve in this way, customers will no longer read these online reviews. We urge the key players in these online markets to find a way to increase the cost of manipulation in order to mitigate the manipulation effect. We call for collective thinking within this community, including the technical vendors and business entities, to build a better online system to fight against this practice. The ideal situation would be that online reviews represent the truth, the whole truth, and nothing but the truth about their products. However, unless we can resolve the manipulation issue, online consumers can only get the "partial truth."

## Appendix A. Additional tests for validating the existence of online reviews manipulation

### A.1. Robustness check: Case I

What we did in Table 4 was to regress the current rating on the lag average rating with time-series dimension (for each item, we include 20 reviews from different times). For robustness checking purposes, out of group 1 and group 2, we chose 20 cross-sectional datasets (cross-sectional datasets means including only the reviews of the same order) and run separate regressions at each order level. We excluded the "*Silence*" variable because there is no information difference to the variable "*Silenc*e" when all reviews are from the same order. Results are presented in Table A1. The coefficients of the intercept of group 1 are consistently significantly greater than zero while the coefficients of the *LagAvegrating* are consistently significantly less than 1 for various orders (orders 6 to 24th). However, for group 2, from order 80 to order 98, their intercepts are zero; while their coefficients of *Lag*(*Avgrating*) are not different from 1 (based on F-test). All the results show that manipulation does exist.

### A.2. Robustness check: Case II

The results in Table 4 might be problematic if the error term is not constant over time. Hu et al. [7] have shown that ratings are more likely to follow U-shaped distributions because consumers are more likely to write reviews when they are very satisfied or dissatisfied. Hence, the error term might not be normal. If as time progresses, customers are more likely to moan, then it might lead to the self-selection documented by [9]. If over time, the relative likelihood of

**Table A1**
Regression result at order level.

| Group 1 | | | | | | | Group 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Order | Coefficient of *intercept* | Coefficient of *Lagavgrating* | Is coefficient of *Lagavgrating* significantly different with 1 | | N | Adj R square | Order | Coefficient of *intercept* | Coefficient of *Lagavgrating* | Is coefficient of *Lagavgrating* significantly different with 1 | | N | Adj R square |
| | | | F-value | Yes/no | | | | | | F-value | Yes/no | | |
| 6 | 1.78*** | 0.58*** | 443.33*** | Yes | 6250 | 0.13 | 80 | 0.03 | 0.98*** | 0.35 | No | 6250 | 0.18 |
| 8 | 1.40*** | 0.64*** | 301.26*** | Yes | 6250 | 0.14 | 82 | −0.05 | 0.96*** | 1.78 | No | 6250 | 0.18 |
| 10 | 1.02*** | 0.74*** | 157.12*** | Yes | 6250 | 0.17 | 84 | −0.03 | 0.99*** | 0.04 | No | 6250 | 0.18 |
| 12 | 0.97*** | 0.75*** | 126.20*** | Yes | 6250 | 0.17 | 86 | −0.07 | 0.96*** | 1.71 | No | 6250 | 0.17 |
| 14 | 0.89*** | 0.75*** | 124.23*** | Yes | 6250 | 0.16 | 88 | −0.7 | 0.99*** | 0.03 | No | 6250 | 0.18 |
| 16 | 1.23*** | 0.71*** | 145.73*** | Yes | 6250 | 0.14 | 90 | 0.01 | 0.98*** | 0.89 | No | 6250 | 0.17 |
| 18 | 0.73*** | 0.78*** | 81.86*** | Yes | 6250 | 0.16 | 92 | 0.01 | 0.98*** | 0.33 | No | 6250 | 0.18 |
| 20 | 0.77*** | 0.78*** | 82.26*** | Yes | 6250 | 0.16 | 94 | −0.05 | 0.99*** | 0.10 | No | 6250 | 0.18 |
| 22 | 0.84*** | 0.76*** | 94.44*** | Yes | 6250 | 0.14 | 96 | 0.21 | 0.96*** | 2.19 | No | 6250 | 0.17 |
| 24 | 0.90*** | 0.74*** | 124.73*** | Yes | 6250 | 0.14 | 98 | −0.08 | 0.96*** | 1.71 | No | 6250 | 0.17 |

*** $P \leq 0.01$.

brag or moan stays the same, then our results should still hold. Furthermore, if the probability of brag or moan changes over time, then results based on OLS estimation might not be valid due to the variance of the error term changing over time. We conduct a White test to check the heteroscedasticity. The results show that heteroscedasticity does exist. As discussed before, we corrected the potential heteroscedasticity in our data according to the method proposed in Long and Ervin [10]. Qualitatively the results still hold.

### *A.3. Robustness check: Case III*

Our results may be driven by the number of the ratings which was used to calculate the *Lag*(*avgrating*). Recall that for each item, group 1 includes the 6th to the 25th review received by each item, while group 2 includes the 81th to the 100th review received by each item. Hence, the *Lag*(*avgrating*) in group 1 (due to the small number of ratings) may include more statistical error and less stability than that of group 2. For example, for review 16 in group 1, we used ratings of the 1st to the 15th review (15 reviews) to estimate its *Lag*(*avgrating*), however, for review 91 in group 2, we used ratings of the 1st to the 90th review (90 reviews) to estimate its *Lag*(*avgrating*). In order to get a comparable estimation of the *Lagavgrating* in these two groups, we defined a new way to estimate *Lagavgrating* for group 2, which uses only the nearest 10 lag reviews to calculate the lag average rating. For example, when the dependent variable *Rating* is the 100th review (order), we just use the mean rating of the 90th review to the 99th review to approximate its independent variable *Lag*(*avgrating*). Qualitatively the regression results still do not change.

## References


[1] W. Antweiler, M.Z. Frank, Is all that talk just noise? The information content of internet stock message boards, The Journal of Finance 59 (3) (2005) 1259–1294.
[2] C. Dellarocas, Strategic manipulation of internet opinion forums: implications for consumers and firms, Management Science 52 (10) (2006).
[3] P. Chatterjee, Online reviews: do consumers use them? Advances in Consumer Research 28 (1) (2001) 129–133.
[4] J. Chevalier, A. Goolsbee, Measuring prices and price competition online: Amazon and Barnes and Noble, Quantitative Marketing and Economics 1 (2) (2003) 203–222.
[5] J. Chevalier, D. Mayzlin, The effect of word of mouth on sales: online book reviews, Journal of Marketing Research 43 (3) (2006) 345–354.
[6] D. Godes, D. Mayzlin, Using online conversations to study word of mouth communication, Marketing Science 23 (4) (2004) 545–560.
[7] N. Hu, P.A. Pavlou, J. Zhang, Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication, ACM Conference on Electronic Commerce (2006) 324–330.
[8] N. Hu, L. Liu, J. Zhang, Do online reviews affect product sales? The role of reviewer characteristics and temporal effects, Information Technology and Management 9 (3) (2008).
[9] X. Li, L.M. Hitt Self, selection and information role of online product reviews, Information Systems Research 19 (2008) 456–474.
[10] J.S. Long, L.H. Ervin, Using heteroscedasticity consistent standard errors in the linear regression model, The American Statistician 54 (2000) 217–224.
[11] D. Mayzlin, Promotional chat on the internet, Marking Science 25 (2) (2006) 157–165.
[12] P.D. Wysocki, Cheap talk on the web: the determinants of postings on stock message boards. Working paper, University of Michigan, 1999.



**Nan Hu** is an Assistant Professor of Accounting and Finance at the University of Wisconsin at Eau Claire. He is also an Assistant Professor of Information Systems at Singapore Management University. He received his Ph.D. from the University of Texas at Dallas. Nan's research focuses on investigating the value implications and market efficiency of both traditional information (e.g. company financial report, analyst forecast, corporate governance, etc.) and non-traditional information (e.g. blog opinion, online consumer reviews, etc.), using a combination of theories from accounting, finance, marketing, information economics, sociology, psychology, and computer science. Nan's research has been published in *MISQ (MIS Quarterly)*, *IEEE Transactions on Engineering Management (IEEE-TEM)*, *JMIS* (*Journal of Management Information Systems*), *CACM* (*Communications of the ACM*), *JCS* (*Journal of Computer Security*), and *IT&M* (*Information Technology and Management*).



**Ling Liu** is an Assistant Professor of Accounting and Finance at the University of Wisconsin at Eau Claire. She received her Ph.D. in Accounting from the University of Texas at Dallas. Her research focuses on market efficiency, corporate governance, and relative performance evaluation. Her research has been published in Decision Support Systems, IEEE Transactions on Engineering Management (IEEE-TMC), Information Technology and Management, and International Journal of Accounting and Information Management.



**Vallabh Sambamurthy** is the Eli Broad Professor of Information Technology at the Eli Broad College of Business at Michigan State University. He served as the Executive Director of the Center for Leadership of the Digital Enterprise between 2004 and 2009. He has previously served on the faculties of the business schools at The University of Maryland and The Florida State University. He has expertise in how firms leverage information technologies in their business strategies, products, services, and organizational processes. His work has been funded by the Financial Executives Research Foundation, the Advanced Practices Council (APC), and the National Science Foundation. His work has been published in journals such as the MIS Quarterly, Information Systems Research, Decision Sciences, Management Science, Organization Science, and the IEEE Transactions on Engineering Management.