The Relationship Between Disclosing Purchase Information and Reputation Systems in Electronic Markets

Completed Research Paper

Marios Kokkodis
Boston College
kokkodis@bc.edu

Theodoros Lappas
Stevens Institute of Technology
tlappas@stevens.edu

Abstract

In this work we investigate how the introduction of the Verified Purchase (VP) badge on Amazon.com affected both the review helpfulness and the product ratings. We first conduct a propensity score matching study and find that all else equal, camera reviews are on average ranked 7 positions higher than non-VP reviews, while book VP reviews are on average ranked 11 positions higher than non-VP reviews. Next, we use a natural experiment setting to identify whether the entry of the VP feature had an effect on the (1) overall review helpfulness (both VP and non-VP reviews), and (2) average product rating. Our results show that the introduction of VP caused an increase in review helpfulness of 7.7% for books, and 1.7% for electronics. Furthermore, it caused on average an increase of 20 and 18 positions in the ranks on book and electronic products respectively.

Keywords: Econometric analyses, Natural language processing, Text mining, Experimental economics, Data analysis, Data mining, Machine learning, Electronic markets, Empirical analysis

Introduction

In the past twenty years, the Internet boom has paved the way for the development and rapid growth of electronic marketplaces such as Amazon.com, Zappos.com, ebay.com, and Newegg.com. Amazon.com, stands out as the leading electronic market in the US, and one of the most valuable brands worldwide[1], with more than $74.45 billion in 2013 net sales and a total of 237 million active customer accounts.

In order to increase the efficiency of their marketplace, these platforms have developed reputation systems that are based on user-submitted product ratings and reviews. These reputation systems have grown to become the main source of information for interested consumers by partially reducing various information asymmetries (Akerlof 1970). Even further, previous research has repeatedly verified the strong economic impact of product ratings and reviews on the marketplace (Archak et al. 2011; Chevalier and Mayzlin 2006; Ghose and Ipeirotis 2011; Li and Hitt 2008). Because of the strong impact of these reputation systems electronic marketplaces have developed mechanisms to identify the most important/high-quality reviews for the consumers. The most commonly used mechanism is peer reviewing, with customers giving helpful votes to reviews in order to signal their quality. The marketplace then incorporates this information into a ranking system that highlights the most helpful reviews.

While helpful votes have been established as the standard way to evaluate reviews and reviewers, marketplaces have experimented with additional features that are meant to serve as proxies of a reviewer's credibility. A characteristic example of such a feature is the introduction of the "Verified Purchase" (VP) badge on Amazon.com. This VP badge appears only next to the reviews of users who have bought the reviewed product from Amazon.com (see Figure 1). On par with Amazon.com, other marketplaces such as Expedia.com and Orbitz.com have since introduced similar badges ("expedia guests" and "verified customers" respectively). The premise of the VP badge is simple: customers who have bought and have first-hand experience with a product are in a better position to evaluate its true quality. Therefore, their reviews are more likely to be objective and be trusted by potential buyers.

Despite the fact that this feature has been in place for a while (Amazon introduced the VP badge in 2009), its effect on the marketplace reputation ecosystem has yet to be documented. Our work is the first to address this task, which we approach from two different vantage points. First, we study whether VP reviews are perceived as more helpful than non-VP ones. We start by collecting Amazon reviews for two types of products: books and electronics. We then analyze this data and extract textual and non-textual features that have been shown in the past to be correlated with review helpfulness. Based on these features, we build a propensity score model for estimating the probability of a review to be VP, and based on this model, we create a matching sample of VP and non-VP reviews. Our results show a strong and positive effect on review helpfulness: all else equal, camera VP reviews are ranked on average 7 positions higher than non-VP reviews, while book VP reviews are ranked on average 11 positions higher than non-VP reviews.

The second part of our study examines the effect of the VP badge on the (1) helpfulness of both VP and non-VP reviews and (2) product rating. Our goal is to identify a causal relationship between the introduction of the VP badge and these two dimensions of reputation. To do so, we exploit a natural experiment setting enabled by the global dimension of Amazon. In 2009, Amazon introduced the VP badge only on its US website. The feature was not introduced to the UK platform (Amazon.co.uk) until 2 years later. In addition to having overlapping product inventories, the two websites use the same product codes for their common products. This allows us to perform a diff-in-diff study on the effect of the VP entry on both the overall review helpfulness and the product rating (i.e., same product, before and after the VP entry in the US, in both markets US and UK). Our results show that the introduction of VP caused an increase in review helpfulness of 7.7% for books, and an increase of 1.7% for electronics, independent of whether or not these reviews carried the VP badge. Furthermore, it caused on average an increase of 20 positions in the ranks on book products, and an increase of 18 positions in the ranks for electronics.

[1] http://www.statista.com/topics/846/amazon/

Our work contributes to the extended literature on online reviews and reputation systems in electronic markets by establishing (1) a positive link between the disclosure of purchase information and the perceived helpfulness of reviews and (2) a product-type dependent link between disclosing purchase information and the overall product rating. The revealing of these relationships provides very important design insights for new (or current) electronic marketplaces. Finally, our study contributes on the methodological frontier, by drawing on multiple disciplines and combining ideas and algorithms from machine learning, text mining, sentiment analysis, social sciences and econometrics.

Background

Electronic word of mouth has been studied from multiple vantage points in the past fifteen years. In this section, we discuss the importance of reputation systems and online reviews in the electronic marketplace ecosystem, while we further connect previous works with our study and highlight our contributions. We provide the necessary background for our study by drawing from four main streams of related research: reputation mechanisms and their importance, the economic impact of product reviews, the trustfulness and manipulation of reviews, and finally, the characteristics and the importance of review helpfulness.

Reputation Mechanisms

Our main goal is to estimate how the disclosure of purchasing information affects the reputation ecosystem of an electronic marketplace. A long line of previous work has established the multidimensional importance of reputation mechanisms in these marketplaces. In particular, researchers have found that reputation mechanisms in electronic markets (1) resolve various information asymmetries (Dellarocas 2003; Dellarocas 2006; Kokkodis and Ipeirotis 2015) and (2) improve transaction efficacy (Bakos and Dellarocas 2011; Bolton et al. 2004). Furthermore, other studies found that (1) negative ratings are far more influential and detrimental than positive ones (Chevalier and Mayzlin 2006; Standifird 2001), (2) reputable sellers create an increase in willingness to pay (Resnick et al. 2006), and (3) reputation scores appear to form J-shaped distributions (Hu et al. 2009). Finally, researches showed a U-shaped relationship between reviewers' posting propensity and their expected product quality (Ho et al. 2014). These findings set the background of our work, and they underline the importance of reputation mechanisms towards increasing efficiency of electronic marketplaces. Our work extends this literature by studying how the introduction of a unique characteristic (VP badge) has affected the complete reputation system (review helpfulness and product ratings) on the largest online market, Amazon.com.

Economic Impact of Online Reviews

On par with the long line of research on reputation systems, a lot of work has focused on the economic impact of online reviews in electronic markets. Specifically, previous works have established that (1) an improvement in ratings leads to an increase in sales (Chevalier and Mayzlin 2006; Chintagunta et al. 2010; Zhou and Duan 2010), (2) both external and internal reviewing platforms have a strong effect on sales (Gu et al. 2012; Lu et al. 2013) and (3) that the impact of reviews on sales decreases with time (Hu et al. 2008). Even further, researchers have proven the relationship of various online review characteristics and (1) the upstream competition between firms (Kwark et al. 2014), (2) different marketing strategies for engaging early product adopters (Li and Hitt 2008) and (3) product growth (Clemons et al. 2006). Finally, a stream of work has focused on building models to predict/explain product sales (Archak et al. 2011; Dellarocas et al. 2007; Ho et al. 2014; Zhang et al. 2010). Given this strong economic impact of online reviews and product ratings, understanding the interplay between different characteristics and the reviewing ecosystem becomes crucial. Our study contributes towards that direction, by studying how the introduction of a new feature affects both the perceived helpfulness of a review but also the review product rating.

Trustfulness and Fraud detection

Since the treatment we examine in this work (i.e., the introduction of the verified purchase badge discussed at the end of this section) is strictly associated with the perceived trustfulness of a review, we briefly mention here previous research on review believability for completeness. A line of work focuses on building models that estimate the review trustfulness (Hu et al. 2006; Kokkodis 2012). Next, researchers have observed that (1) identity-relevant reviewer information (Forman et al. 2008), (2) argument quality (Cheung et al. 2012) and (3) reviewers' popularity (Goes et al. 2014) increase review believability. Finally, a previous paper by (Hu et al. 2011) has shown the existence of a monotonically decreasing relationship between the manipulation of reviews and the product's true quality .

Review Helpfulness

Disclosing purchase information can be seen as a review characteristic that might have a direct impact of the perceived helpfulness. As a result, our work relates to a series of previous studies that focused on extrapolating relationships between various review features and the perceived review helpfulness. From a marketplace perspective, having helpful reviews increases the platform's welfare by facilitating product selection. Throughout the years, academics have established links between the review helpfulness and the (1) review length (Kim et al. 2006; Wu et al. 2011), (2) review text (Kim et al. 2006; Liu et al. 2008), (3) review believability and objectivity (Ghose and Ipeirotis 2011; Otterbacher 2009), (4) product rating (Danescu-Niculescu-Mizil et al. 2009; Kim et al. 2006; Mudambi and Schuff 2010), (5) reviewer's history (Ghose and Ipeirotis 2011; Liu et al. 2008), (6) reviewer's identity and social network (Lu et al. 2010), (7) product type (Mudambi and Schuff 2010), (8) emotions (anxiety/anger) (Yin et al. 2014) and (9) readability (Ghose and Ipeirotis 2011; Wu et al. 2011). One step further, researchers have proposed algorithms for identifying a comprehensive and diversified set of reviews (Lappas and Gunopulos 2010; Tsaparas et al. 2011). Our study extends the previous works on helpfulness by examining its relationship with the disclosure of purchase information. Even further, we use this long line of research as a guideline for selecting our set of control variables. We discuss this in more detail in the later sections.

Contribution of our work

Our work extends the current literature on online reviews and reputation systems in electronic markets in the following ways: First, we find a positive relationship between the disclosure of purchase information and the perceived helpfulness of a review. Second, by exploiting a natural experiment setting, we establish the existence of a positive link between revealing purchase information and the overall review helpfulness of the platform. Third, we find a product-type dependent relationship between disclosing purchase information and the overall product rating. Furthermore, given the established importance of review helpfulness and product ratings and their associations with market efficiency and product sales, our study reveals important design insights for new (or current) electronic marketplaces. Finally, from a methodological perspective, our study provides a technically solid framework for combining methodologies from machine learning text mining and sentiment analysis, to social sciences and econometrics.

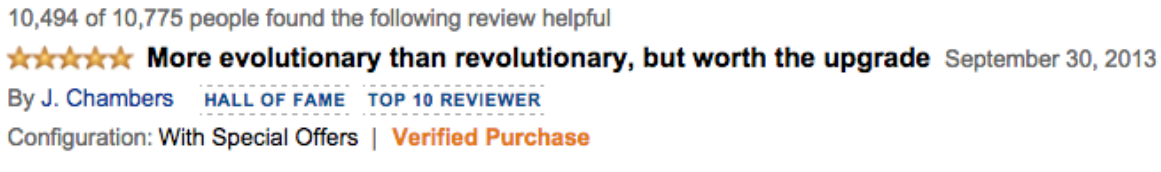

Figure 1: Example of a VP Amazon.com review header

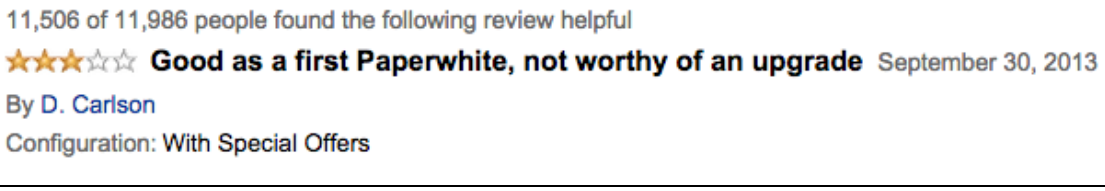

Figure 2: Example of a non-VP Amazon.com review header

The Introduction of Verified Purchases

In September 2009[2] Amazon.com introduced the "verified purchase" feature[3]. Taken from the Amazon.com announcement: "*When a product review is marked "Amazon Verified Purchase," it means that the customer who wrote the review purchased the item at Amazon.com […] Customers reading an Amazon Verified Purchase review can use this information to help them decide which reviews are most helpful in their purchasing decisions.*"

The announcement makes it clear that Amazon expects the introduction of this feature to boost the review helpfulness within its marketplace. In Figure 1 we show the header of a VP review. This review has the Verified Purchase (VP) badge, along with other reviewer badges such as "Hall of Fame" and "Top 10 Reviewer". To the contrary, in Figure 2, we show a review that does not have the VP badge.

As we mentioned in the introduction, the focus of this study is twofold. First, we are interested whether reviews that carry the VP badge are all else equal more helpful than reviews that do not carry the VP badge. Second, we want to understand how the entry of the VP feature affects the overall helpfulness of the entire population of reviews as well as the overall product rating. We discuss these questions next.

Are VP Reviews More Helpful than non-VP?

In this section, we focus on studying whether verified reviews are more helpful than the non-verified ones. Specifically, we conduct a propensity score matching study (Austin 2011) that controls for a series of confounding factors that have been found in the past to be associated with review helpfulness. We define helpfulness – in accordance with a series of previous studies (Danescu-Niculescu-Mizil et al. 2009; Ghose and Ipeirotis 2011; Kim et al. 2006) – as the ratio between helpful votes and the number of total votes:

$$helpfulness = \frac{helpful\ votes}{total\ votes}$$

In the rest of this section we describe our methodology and the datasets that we used, followed by our findings.

[2] http://goo.gl/kmmahj

[3] http://goo.gl/5L8H5Q

Data[4]

To perform our empirical analysis, we collect reviews from Amazon.com, for two product categories: Books and Electronics. Books are experience products (Bei et al. 2004; Nelson 1970), i.e., the true quality of a book is highly subjective and is revealed to its buyer post-purchase. On the other hand electronics have standard technical characteristics that can be reviewed/criticized objectively and according to the market benchmarks. Because of the very different characteristics of these two product categories, we expect the introduction of the VP-badge to have category-dependent effect on review helpfulness.

Our electronics dataset consists of a total of 190,000 reviews on 3037 products, while our books dataset consists of 183,000 reviews on 3762 books. These datasets are considerably larger than those used in most of the previous related studies, e.g., (Ghose and Ipeirotis 2011; Mudambi and Schuff 2010). To create these datasets we scraped product codes from Amazon.com searchers on "books" and "electronics", and then visited each product page to collect the posted reviews. Our final set of reviews spans eighteen years (since 1998).

In **Table 1** we present additional statistics about these two datasets, including information about the helpfulness and the number of VP reviews. One interesting observation is that the number of verified reviews in electronics is significantly larger than the respective number in books. The reason is that the technological evolution in electronics (electronics) is high paced: electronics that are cutting edge today usually become obsolete in two or three years. As a result, our crawling task collects relatively new reviews on electronics (i.e., when the VP badge was already in place). On the other hand, popular books tend to remain popular for much longer periods of time. Hence many of the collected book reviews were posted long before the VP introduction.

Table 1: Propensity score study dataset statistics

Dataset	AVG(Helpfulness)	St. Dev.	# VP reviews	# reviews	# Products
Electronics	0.862	0.285	139530	190,020	3037
Books	0.816	0.283	69659	183585	3762

Finally, in addition to these two datasets, we further used a separate set of 40,000 reviews (20,000 with product rating 5 and 20,000 with product rating 1) across both books and electronic products as a training set to build our sentiment analysis models (described in the next subsection). We will refer to this dataset as the "sentiment" dataset.

Methodology

To study the effect of the VP badge on review helpfulness, we perform a propensity score matching (PSM) study. Propensity score techniques provide us with the tools to design and analyze an observational study that mimics some of the particular characteristics of a randomized trial. In short, propensity score studies reduce or eliminate the effect of confounding variables in the presence of observational data. Multiple previous works have used propensity score methods to study (1) the distinction between influence-based contagion and homophilly-driven diffusion (Aral et al. 2009), (2) the effects of kindergarden retention on children's social-emotional development (Hong and Yu 2008), (3) the effectiveness of alcoholics anonymous (Ye and Kaskutas 2009), and (4) the effects of small school size on mathematics achievement (Wyse et al. 2008).

The input to PSM consists of a population of individuals with various attributes, a treatment that is associated with a subset of the population, and a selected attribute of interest α^*. First, a logistic regression is performed to estimate the probability that each member of the population is treated. This

[4] The datasets (and the code – Python and Java) used in this study can be made available for research purposes. Please contact the authors if interested.

probability is referred to as the member's propensity score. The treatment serves as the (binary) dependent variable for the regression, while the vector of covariates Z includes all of the population's attributes except α*. Formally:

$$Propensity\ Score\ of\ i^{th}\ review := \Pr(treated \mid Z_i) = \Pr(VP_i = 1 \mid Z_i) = \frac{\exp(Z_i)}{1 + \exp(Z_i)}$$

The computed probabilities are then used to create pairs of treated and untreated individuals (Aral et al. 2009): a treated individual *i* is matched with an untreated individual *j* if:

$$\min \left|\Pr(i = \text{Treated}|Z_i) - \Pr(j = \text{Treated}|Z_j)\right| < 2\sigma_d, \qquad (1)$$

, where $d = \Pr(i = \text{Treated}|Z_i) - \Pr(i = \text{Treated}|Z_j)$. The α* values of the two individuals in a pair can then be confidently compared without concern for the effects of the attributes in Z. In our context, the attribute of interest α* is the helpfulness of the review, while the treatment is whether or not the review carries the VP badge. We describe the vector of covariates Z next.

Feature Vector Z

Our feature vector *Z* includes attributes that have been found in the past to be correlated with review helpfulness. In particular, we start by extracting information from the review text. We create a feature vector of unigrams (Manning and Schütze 1999), which have been previously found to be closely associated with the helpfulness of a review (Kim et al. 2006). The unigram vectors for books and electronics consist of 1000 word-features each. Examples of word-features include "addition", "appreciate", "brilliant", etc.

Beyond the actual text, the product rating of the review has also been found to be correlated with the perceived helpfulness of the review (Kim et al. 2006; Mudambi and Schuff 2010). Specifically, prior research showed that review extremity in combination with the product type has a strong effect on perceived helpfulness (Mudambi and Schuff 2010). To control for review extremity, we include in our feature vector *Z* the actual product rating of the review (1 to 5 star rating).

Furthermore, the sentiment of the review has been proven to be correlated with its helpfulness (Ghose and Ipeirotis 2011). To include information about the sentiment of each review, we build probabilistic models that estimate the probability of a review to be of positive sentiment – similar to (Pang and Lee 2008). In particular, we use the sentiment dataset described before as training set and build three different probabilistic models: a language model, a naïve Bayes, and a logistic regression model. The selection of these models is not random, since all of them have been shown to perform extremely well in text classification tasks (Ifrim et al. 2008; Manning and Schütze 1999; McCallum and Nigam 1998)

For all three models we run a ten-folded cross validation on the training set. We present the results in **Table 2**. The Naïve Bayes (NB) classifier performs with an Accuracy of 84.7%, and with an AUC score of 0.928. Since the training set is balanced across positive and negative instances, these values represent great performance. Our language model, which we implemented by using the Lingpipe library[5] in Java, shows an improved performance over the NB approach (with an accuracy of 0.86 and AUC of 0.92). Finally, the performance of our logistic regression model significantly outperforms both the NB and the language model approaches, in both evaluation metrics. As a result, we choose to use the logistic regression model to estimate the probability of a review to have positive sentiment. In particular, we build the logistic model on the "sentiment" dataset, and use it to estimate the probably of a review to have positive sentiment on each instance of our electronics and books datasets.

[5] http://alias-i.com/lingpipe/index.html

Table 2: Sentiment Analysis Classifiers		
Model	**ACC**	**AUC**
Naïve Bayes	0.847	0.928
Language Model	0.862	0.920
Logistic Regression	0.869	0.939

Table 3: Logistic Regression 10-fold cross validation Results					
Dataset	ACC	AUC	Min Propensity Score	Max Propensity Score	St. Dev. (Propensity Score)
Electronics	0.793	0.789	0.000	0.996	0.219
Books	0.761	0.815	0.000	0.966	0.258

Finally, for additional control variables, we further include the total number of votes of the review, as well as personal information about the reviewer (i.e., whether or not the reviewer has an Amazon profile, and whether or not the reviewer has any badges[6], similar to (Ghose and Ipeirotis 2011)). The final version of vector *Z* includes 1008 variables. By controlling for all these confounding factors we achieve the isolation of the effect of carrying the VP badge on the perceived helpfulness.

Findings

The ten-folded cross-validation results of our propensity score logistic regression are shown in Table 3. We are mostly interested in the AUC values, which represent the probability of correctly ranking a positive instance (i.e., a VP review) over a negative one (i.e., a non-VP review) (Provost and Fawcett 2001). From the table, we see that these values are 78.9% for electronics, and 81.5% for books, signifying a very good performance of our model for predicting whether or not a review has been treated.

After matching treated with non-treated reviews with similar propensity scores (according to Equation 1) we end up with 86,280 matched book reviews, and 76,012 matched camera reviews. In Figure 3, we show the distribution of helpfulness for the two types of products in our study, for the treated and the non-treated instances in our matching samples. We observe that, in both graphs, the treated distributions are significantly skewed to the right (higher helpfulness values) compared to the not treated ones. These graphs are the first observable evidence that consumers perceive VP reviews as more helpful than the non-VP ones.

[6] http://goo.gl/cxolRQ

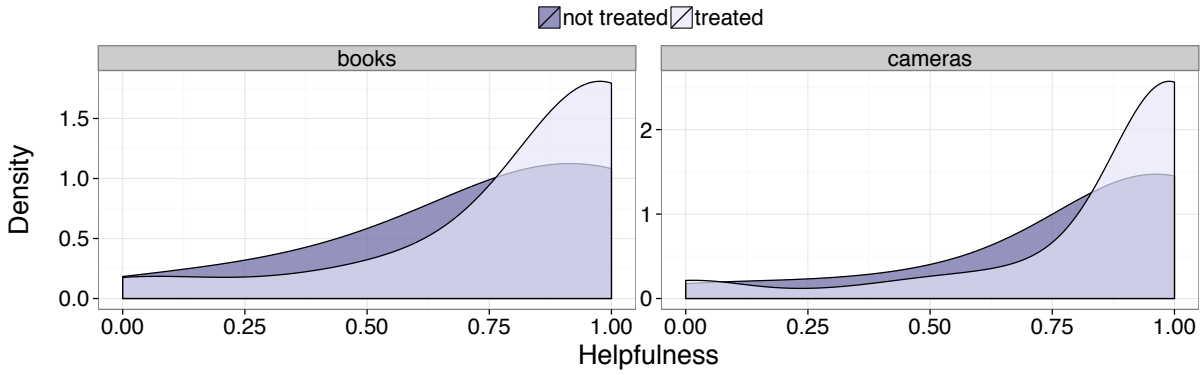

Figure 3: The distributions of helpfulness for treated and not treated instances in our matched samples, for the two product categories we consider.

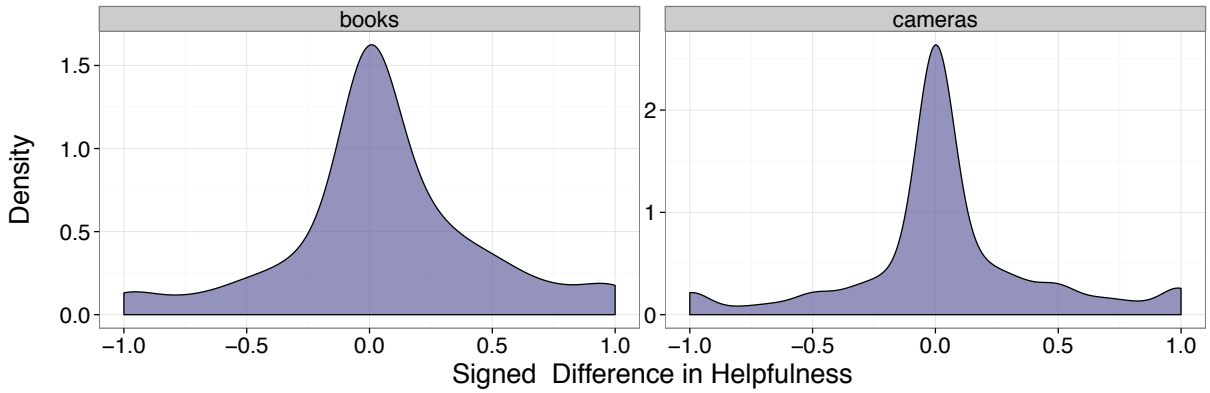

Figure 4: The distributions of difference in helpfulness between our matched samples for the two product categories we consider.

To get a better picture of the effect, we further compute the difference in terms of helpfulness between each matched pair. The distributions of the differences are shown in Figure 4. We observe that the right side of the distributions (positive difference) contains a larger probability mass than the left side, in both datasets.

Finally, to quantify these observations, we compute the descriptive statistics of these two distributions of differences. We present the results in Table 4. The mean difference between treated and not treated reviews is 0.036 for electronics and 0.056 for books. In other words, on average, all else equal, (1) a VP review in books is expected to be 0.056 more helpful than a non-VP review; (2) a VP review in electronics

is expected to be 0.036 more helpful than a non VP review. These differences might appear tiny at first, but they represent a boost in the review rankings of 11 and 7 positions respectively!

To check whether these results are significant, we run an one sample t-test, where:

$$H_0: \mu = 0$$

$$H_1: \mu > 0$$

The zero p-values shown in Table 4 suggest that we should reject the null hypothesis.

Table 4. : Difference in helpfulness between matched instances

Dataset	Mean	p-value
Electronics	0.036	0.000
Books	0.056	0.000

To conclude, in this section we performed a propensity score matching study to quantify the effect of VP badge on the helpfulness of the review. We found that reviews that carry the verified badge are significantly more helpful than the ones that do not. Next, we focus on quantifying the effect of the VP introduction on the complete reviewing ecosystem (i.e., helpfulness and product ratings).

The Effect of the VP Entry on the Reviewing Ecosystem

Our first study showed that VP reviews are all else equal more helpful than non-VP ones. However, the introduction of the Verified Purchase badge could have had an effect on the review population as a whole (i.e., both VP and non-VP reviews). For example, reviewers with intention to create fake reviews might be discouraged by the fact that (1) their reviews might lose credibility when they are non-VP and/or (2) they will have to buy the product and then write the review. In other words, the VP feature could act as an additional barrier for creating low-quality or misleading reviews, and as a result, we would expect to see an overall rise of review helpfulness, independent of whether or not the reviews are VP.

The global expansion of Amazon in combination with the time variability of the VP introduction in different markets creates a natural experiment setting that facilitates the study of our question. As we mentioned earlier, the verified purchase attribute was introduced on Amazon.com (US) on September 18th 2009. During the same period, Amazon.co.uk (the UK version of Amazon) was running without the feature, which introduced later, sometime in March 2012[7]. Because Amazon uses the same unique product identifiers across its platforms globally (i.e., ASIN ids), we have the ability to study how the set of reviews for the same products evolved in these different platforms, before and after the introduction of VP in the US, and during the same time periods in the UK.

[7] http://www.amazon.co.uk/forum/top%20reviewers?_encoding=UTF8&cdForum=Fx2AH5S1CY4QEMR&cdThread=Tx20M1IH9BVFUK7

Data

To perform this study we need a different dataset than the one we used before. Specifically, we need to find products that used to be available in both the UK and the US markets before and after the VP entry in the US (2009). Even then, in order to acquire low-variance estimates of the average helpfulness (D.V. of interest) of a set of product reviews, we will need these products to have a critical mass of reviews in both time periods and in both markets. These requirements are far from trivial for mainly two reasons: very few products that used to be available earlier than 2009 still exist and (2) the distribution of number of reviews written per day is negatively correlated with time, which means that products that were available on Amazon a few years before 2009 would be highly unlikely to have been reviewed after 2009 (see Figure 5). Furthermore, in order to create the two periods of interest (before and after the VP introduction in the US) we consider only reviews written between 2008-01-01 and 2010-12-31.

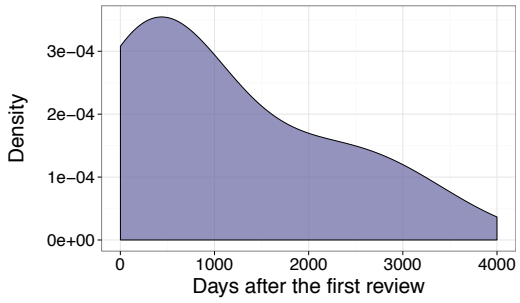

Figure 5: When people post product reviews on Amazon.com

Because of these peculiarities, and in order to create a big enough dataset of products that meet these requirements, we ran a set of multiple parallel massive crawling tasks on Amazon.com and Amazon.co.uk, which resulted in 7.3 million reviews across 1.7 million products in our two categories, books and electronics. Note also that our crawlers were targeting products that used to be available in 2011 and 2012 in Amazon.com (i.e., closer to our dates of interest)[8].

After collecting all these reviews we were able to create a balanced set of 266 electronic products, and a balanced set of 2035 books. These two final datasets include products that have at least three reviews in both time periods and both markets.

Methodology

Similarly to our propensity study, we use findings from earlier studies to build the set of control variables that are associated with review helpfulness. The main difference in this analysis is that we are now interested in the average helpfulness of a set of reviews on the product level. This peculiarity poses restrictions in the use of a bag of words approach (i.e., we should accumulate all the reviews within a product) and as a result we do not include pure review text in our modeling.

[8] Part of these set of products was targeted through an older Amazon dataset used in (McAuley and Leskovec 2013).

The set of dependent and independent variables in this study include the average helpfulness, the average product rating, the average sentiment of the reviews within a product (which we estimate as described in the previous section), the average length of the review and the average signed difference between the product rating of a review and the average product rating – which has been found to be correlated with the helpfulness and the product rating of the review (Danescu-Niculescu-Mizil et al. 2009).

As we mentioned before, we split our data into two periods, in order to employ a diff-in-diff analysis. The first period (between 01-01-2008 and 09-18-2009) includes product reviews before the appearance of the Verified Purchase feature, in both the US and the UK marketplaces. The second period (between 09-18-2009 and 12-31-2010) includes reviews after the VP introduction in the US, but before the VP introduction in the UK.

The descriptive statistics of all the variables in our dataset are shown in Table 5. The correlations between all the variables are shown in Figure 6. We observe that there is not a single pair of variables that correlate higher than 50%.

Table 5: Summary Statistics

	Variable	**Obs.**	**Mean**	**Std. Dev.**	**Min**	**Max**
Books	AVG(Helpfulness)	17256	0.8	0.18	0	1
	AVG(Product Rating)	17256	4.13	0.68	1	5
	AVG(Signed Diff)	17256	-0.01	0.33	-2.25	2.47
	AVG(Sentiment)	17256	0.69	0.35	0	1
	AVG(Length)	17256	183.5	99.7	19.8	1581.4
	VP	17256	0.25	0.43	0	1
Electronics	AVG(Helpfulness)	2261	0.88	0.11	0.33	1
	AVG(Product Rating)	2261	4.19	0.62	1	5
	AVG(Signed Diff)	2261	-0.01	0.32	-2.0	1.83
	AVG(Sentiment)	2261	0.56	0.33	0	1
	AVG(Length)	2261	103.7	73.1	1	563.4
	VP	2261	0.25	0.43	0	1

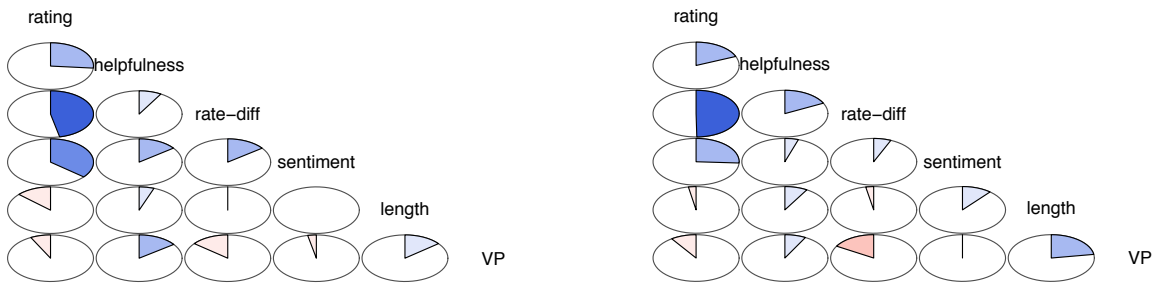

Figure 6: The correlograms for our set of variables, for books (left) and electronics (right). Negative correlations are formed counter-clock wise and with a red hue.

The effect of the introduction of VP on the overall review helpfulness

To study the effect of the introduction of the Verified Purchase feature on the average review (VP and Non-VP) helpfulness we estimate the following model:

$$\log\big(avg(helpfulness)\big) = A_p + B_m + C_t + \beta_1.avg(productRating) + \beta_2.avg(signedDiff)$$
$$+\beta_3.avg(sentiment) + \beta_4.avg(length) + \gamma.VP_{mt} + e_{pmt}$$

In this model, A_p B_m and B_m represent the fixed effects of the products, markets (US,UK) and time (for every year in our study). The coefficient of interest, γ, is the difference-in-difference estimate of the effect of the VP entry on helpfulness. If $\gamma > 0$, then the introduction of VP caused a positive effect on the average helpfulness of reviews, independent of whether or not the reviewer has bought the product from amazon (VP vs. non-VP reviews).

In **Table 6** we show the results of our fixed effects (FE) models for books and electronics. The variable of interest is the entry of verified purchase on Amazon.com (VP). We observe that in both product types, the coefficient of VP is positive and significant. In books, the entry of VP caused a 7.7% increase on helpfulness across all reviews (verified and non-verified ones), while in electronics it caused an increase of 1.7%.

The R-squared values of our models range between 0.06 and 0.1 across the two product categories. These R-squared values are for the "within" (differenced) fixed effect estimators that estimate this regression by differencing out the average values across products. The R-squared reported is obtained by only fitting a mean deviated model where we assume the effects of the groups to be fixed quantities. As a result, all of the group effects are simply subtracted out of the model and no attempt is made to quantify their overall effect on the fit of the model. Hence, the calculated R-squared values do not take into account the explanatory power of the fixed effects. This is also consistent with the work of (Ghose and Ipeirotis 2011) on review helpfulness.

Table 6: Effect of VP entry on Books and Electronics. Significance codes: "*" 0.001, "**" 0.01, "*" 0.05, "." 0.1**		
	Books	Electronics
DV: Log(AVG(Helpfulness))	FE	FE
AVG(Product Rating)	0.077***	0.033**
AVG(Signed Diff)	-0.011	0.014
AVG(Sentiment)	0.012***	-0.005.
AVG(Length)	0.057***	0.02***
VP	**0.077*** **	**0.017***
Observations	17256	2261
R-squared	0.06	0.10

Correctness of our Model Specification

To verify the correctness of our model specification, we run a series of tests. In particular, in order to check whether a random effects model should be estimated, we perform the Hausman test (Greene 2008): we assume that the null hypothesis is that the preferred model is a random effects model – in practice we test whether the errors are correlated with the regressors. Our test rejects the null hypothesis with p-value of 0.000 in books, and with p-value of 0.002 in electronics. Hence our fixed effects specification is the most appropriate.

Next, we test for serial correlation – errors that are correlated between the two time periods we consider. To do so, we perform the Breusch-Godfrey/Wooldridge test for serial correlation in panel models. In both our datasets we reject the null (i.e., no serial correlation) with p-values = 0.000.

Finally we test for heteroskedasticity. In particular, we use the Breusch-Pagan test (Greene 2008) and we set the null hypothesis to be the existence of homoscedasticity in our model. We reject the null hypothesis, for both datasets, with p-values=0.000. This indicates that heteroskedasticity exists in our specification. In order to control for this heteroskedasticity, we perform robust covariance matrix estimation. In particular, we use the Arellano estimator, which is usually recommended for fixed effects models (Arellano 1987). In **Table 7** we show the resulting coefficients.

We observe that after controlling for heteroskedasticity, our coefficient of interest (VP) has p-value = 0.055 in electronics. and a p-value = 0 in books.

Endogeneity concerns In this study, we interpreted the results of our specification as if the introduction of VP in the US was exogenous to the review helpfulness. We believe that concerns of endogeneity existence are limited. It is highly unlikely and counterintuitive that Amazon identified an increasing trend in terms of helpfulness, and because of that decided to introduce the VP feature. Hence, we argue that the VP entry was exogenous and it was not driven by changes in helpfulness.

Table 7: Arrelano Estimator Coefficients. Significance codes: "*" 0.001, "**" 0.01, "*" 0.05, "." 0.1**		
	Books	Electronics
	FE	FE
AVG(Product Rating)	0.078***	0.033*
AVG(Signed Diff)	-0.011	0.014*
AVG(Sentiment)	0.012**	-0.005.
AVG(Length)	0.056***	0.020***
VP	0.077***	0.017.

The effect of the introduction of VP on the product rating

Here we will use the same dataset with before to study whether the introduction of the VP badge caused any effect on the overall product ratings. Our specification now becomes:

$$\log(avg(productRating)) = A_p + B_m + C_t + \beta_1.avg(avg(Helpfulness)) + \beta_2.avg(signedDiff)$$
$$+\beta_3.avg(sentiment) + \beta_4.avg(length) + \gamma.VP_{mt} + e_{pmt}$$

Table 8: Effect of VP entry on Product Ratings. Significance codes: "*" 0.001, "**" 0.01, "*" 0.05, "." 0.1**		
	Books	Electronics
DV: Log(AVG(Product Rating))	Fixed Effects	FE
AVG(Helpfulness)	0.269***	0.18***
AVG(Signed Diff)	0.000	-0.007***
AVG(Sentiment)	0.001	0.000
AVG(Length)	0.003***	0.000
VP	**0.005*** **	**0.006** **
Observations	17256	2261
R-squared	0.93	0.96

In this specification, if $\gamma > 0$, then the introduction of VP caused a positive effect on the average product rating, otherwise it caused a negative effect.

In Table 8 we show the results of our fixed effects (FE) models for books and electronics. The variable of interest is the entry of verified purchase on Amazon.com (VP). We observe that the introduction of VP caused a 0.5% increase on the product rating of books, and a 0.6% increase on the product rating of

electronics. These effects appear almost trivial in magnitude, however they represent on average an increase/decrease of 20 positions in the ranks for books[9] and 18 positions for electronics.

Assuming that the introduction of the VP badge increased the objectiveness of the reputation ecosystem of Amazon.com, we can argue that overall, before the VP badge reviewers have been rating both books and electronics had a negative bias.

Table 9: Arrelano Estimator Coefficients.
Significance codes: "*" 0.001, "**" 0.01, "*" 0.05, "." 0.1**

	Books	Electronics
	FE	FE
AVG(Helpfulness)	0.27***	0.18***
AVG(Signed Diff)	-0.000	-0.007
AVG(Sentiment)	0.000	0.000
AVG(Length)	0.003***	0.000
VP	0.005***	0.005**

Correctness of our Model Specification

Similar to our analysis before, we run a series of robustness checks. We start by checking whether a random effects model should be estimated. Our test rejects the null hypothesis with p-value of 0.000 in both categories, books and electronics. Hence our fixed effects specification is the most appropriate.

Next, we test for serial correlation – errors that are correlated between the two time periods we consider. To do so, we perform the Breusch-Godfrey/Wooldridge test for serial correlation in panel models. In both our datasets we reject the null (i.e., no serial correlation) with p-values = 0.000.

Finally we test for heteroskedasticity. In particular, we use the Breusch-Pagan test (Greene 2008) and we set the null hypothesis to be the existence of homoscedasticity in our model. We reject the null hypothesis, for both datasets, with p-values=0.000. This indicates that heteroskedasticity exists in our specification. Similar to our previous section, we use the Arellano estimator. In Table 9 we show the resulting coefficients, which are almost identical with the ones we showed in Table 8.

[9] We compute the position increase by estimating the effect on the average review score:

$$Positions\ Increase = \frac{avg(Product\ Rating) * coefficient}{avg(rate\ difference\ in\ ranks)}$$

Conclusions, Implications and Future Directions

In this work, we studied how the introduction of the Verified Purchase (VP) feature affected both the review helpfulness and the product ratings on the Amazon platform. We first conducted a propensity score matching study and found that all else equal, camera reviews are on average ranked 7 positions higher than non-VP reviews, while book VP reviews are on average ranked 11 positions higher than non-VP reviews. Next, we used a natural experiment setting to study whether the entry of the VP feature had an effect on the review helpfulness of both VP and non-VP reviews, as well as on the average product rating. Our results showed that the introduction of the VP badge facilitated an increase in helpfulness of 7.7% and an increase on product ratings of 20 positions in the ranks in books, and an increase in helpfulness of 1.6% and of 24 positions in the product ranks in electronics.

One limitation of our second study is that we did not include the actual review text in our models that estimate helpfulness. We believe that a text analysis and in particular understanding whether the VP introduction changed the way that people write reviews is a very interesting and hard to tackle question, which we intend to address in the future. Furthermore, due to space limitations, we did not include any falsification tests to establish the robustness of the causal link between the introduction of VP and our two dependent variables – we intend to perform these checks in the future. Regardless of these limitations, we believe that our study, even at its current form provides evidence that the VP entry had a significant effect on both the review helpfulness and the product ratings.

In this study we did not focus on understanding whether or not the introduction of the VP badge had any effect on product sales. This is not possible sine that there is no panel dataset on sales that goes back to and before 2009. However, and given the verified link between product ratings and sales (Chevalier and Mayzlin 2006) we would expect that the VP badge introduction would have an indirect effect (through product ratings) on sales. Since we are not able to verify this, we will not discuss it further.

Finally, our work has a strong managerial implication for electronic markets similar to Amazon.com: we have established the existence of a causal link between disclosing purchase information and (1) the perceived helpfulness of the reviews and (2) the overall product rating. The overall positive effect on helpfulness should encourage platform managers to introduce badges similar to the VP, however the controversial (positive/negative) effect on the product rating should guide managers to run randomized trials on their platforms and focus on the effect of the badge introduction on sales.

References

Akerlof, G. A. 1970. "The Market for" Lemons": Quality Uncertainty and the Market Mechanism," *The quarterly journal of economics*), pp. 488-500.

Aral, S., Muchnik, L., and Sundararajan, A. 2009. "Distinguishing Influence-Based Contagion from Homophily-Driven Diffusion in Dynamic Networks," *Proceedings of the National Academy of Sciences* (106:51), pp. 21544-21549.

Archak, N., Ghose, A., and Ipeirotis, P. G. 2011. "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Science* (57:8), pp. 1485-1509.

Arellano, M. 1987. "Practitioners'corner: Computing Robust Standard Errors for within-Groups Estimators*," *Oxford bulletin of Economics and Statistics* (49:4), pp. 431-434.

Austin, P. C. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivariate behavioral research* (46:3), pp. 399-424.

Bakos, Y., and Dellarocas, C. 2011. "Cooperation without Enforcement? A Comparative Analysis of Litigation and Online Reputation as Quality Assurance Mechanisms," *Management Science* (57:11), pp. 1944-1962.

Bei, L.-T., Chen, E. Y., and Widdows, R. 2004. "Consumers' Online Information Search Behavior and the Phenomenon of Search Vs. Experience Products," *Journal of Family and Economic Issues* (25:4), pp. 449-467.

Bolton, G. E., Katok, E., and Ockenfels, A. 2004. "How Effective Are Electronic Reputation Mechanisms? An Experimental Investigation," *Management science* (50:11), pp. 1587-1602.

Cheung, M., Sia, C.-L., and Kuan, K. K. 2012. "Is This Review Believable? A Study of Factors Affecting the Credibility of Online Consumer Reviews from an Elm Perspective," *Journal of the Association for Information Systems* (13:8), pp. 618-635.

Chevalier, J. A., and Mayzlin, D. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of marketing research* (43:3), pp. 345-354.

Chintagunta, P. K., Gopinath, S., and Venkataraman, S. 2010. "The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation across Local Markets," *Marketing Science* (29:5), pp. 944-957.

Clemons, E. K., Gao, G. G., and Hitt, L. M. 2006. "When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry," *Journal of Management Information Systems* (23:2), pp. 149-171.

Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., and Lee, L. 2009. "How Opinions Are Received by Online Communities: A Case Study on Amazon. Com Helpfulness Votes," *Proceedings of the 18th international conference on World wide web*: ACM, pp. 141-150.

Dellarocas, C. 2003. "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," *Management science* (49:10), pp. 1407-1424.

Dellarocas, C. 2006. "Reputation Mechanisms," *Handbook on Economics and Information Systems*), pp. 629-660.

Dellarocas, C., Zhang, X. M., and Awad, N. F. 2007. "Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures," *Journal of Interactive marketing* (21:4), pp. 23-45.

Forman, C., Ghose, A., and Wiesenfeld, B. 2008. "Examining the Relationship between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets," *Information Systems Research* (19:3), pp. 291-313.

Ghose, A., and Ipeirotis, P. G. 2011. "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," *Knowledge and Data Engineering, IEEE Transactions on* (23:10), pp. 1498-1512.

Goes, P. B., Lin, M., and Au Yeung, C.-m. 2014. "'"Popularity Effect" in User-Generated Content: Evidence from Online Product Reviews," *Information Systems Research* (25:2), pp. 222-238.

Greene, W. H. 2008. *Econometric Analysis*. Granite Hill Publishers.

Gu, B., Park, J., and Konana, P. 2012. "Research Note-the Impact of External Word-of-Mouth Sources on Retailer Sales of High-Involvement Products," *Information Systems Research* (23:1), pp. 182-196.

Ho, Y.-C. C., Wu, J., and Tan, Y. 2014. "Disconfirmation Effect on Online Rating Behavior: A Dynamic Analysis," *Junjie and Tan, Yong, Disconfirmation Effect on Online Rating Behavior: A Dynamic Analysis (October 25, 2014)*).

Hong, G., and Yu, B. 2008. "Effects of Kindergarten Retention on Children's Social-Emotional Development: An Application of Propensity Score Method to Multivariate, Multilevel Data," *Developmental Psychology* (44:2), p. 407.

Hu, N., Liu, L., and Sambamurthy, V. 2011. "Fraud Detection in Online Consumer Reviews," *Decision Support Systems* (50:3), pp. 614-626.

Hu, N., Liu, L., and Zhang, J. J. 2008. "Do Online Reviews Affect Product Sales? The Role of Reviewer Characteristics and Temporal Effects," *Information Technology and Management* (9:3), pp. 201-214.

Hu, N., Pavlou, P. A., and Zhang, J. 2006. "Can Online Reviews Reveal a Product's True Quality?: Empirical Findings and Analytical Modeling of Online Word-of-Mouth Communication," *Proceedings of the 7th ACM conference on Electronic commerce*: ACM, pp. 324-330.

Hu, N., Zhang, J., and Pavlou, P. A. 2009. "Overcoming the J-Shaped Distribution of Product Reviews," *Communications of the ACM* (52:10), pp. 144-147.

Ifrim, G., Bakir, G., and Weikum, G. 2008. "Fast Logistic Regression for Text Categorization with Variable-Length N-Grams," *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 354-362.

Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. 2006. "Automatically Assessing Review Helpfulness," *Proceedings of the 2006 Conference on empirical methods in natural language processing*: Association for Computational Linguistics, pp. 423-430.

Kokkodis, M. 2012. "Learning from Positive and Unlabeled Amazon Reviews: Towards Identifying Trustworthy Reviewers," *Proceedings of the 21st international conference companion on World Wide Web*: ACM, pp. 545-546.

Kokkodis, M., and Ipeirotis, P. G. 2015. "Reputation Transferability in Online Labor Markets," *Management Science* (62:6), pp. 1687-1706.

Kwark, Y., Chen, J., and Raghunathan, S. 2014. "Online Product Reviews: Implications for Retailers and Competing Manufacturers," *Information systems research* (25:1), pp. 93-110.

Lappas, T., and Gunopulos, D. 2010. "Efficient Confident Search in Large Review Corpora," in *Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 195-210.

Li, X., and Hitt, L. M. 2008. "Self-Selection and Information Role of Online Product Reviews," *Information Systems Research* (19:4), pp. 456-474.

Liu, Y., Huang, X., An, A., and Yu, X. 2008. "Modeling and Predicting the Helpfulness of Online Reviews," *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*: IEEE, pp. 443-452.

Lu, X., Ba, S., Huang, L., and Feng, Y. 2013. "Promotional Marketing or Word-of-Mouth? Evidence from Online Restaurant Reviews," *Information Systems Research* (24:3), pp. 596-612.

Lu, Y., Tsaparas, P., Ntoulas, A., and Polanyi, L. 2010. "Exploiting Social Context for Review Quality Prediction," *Proceedings of the 19th international conference on World wide web*: ACM, pp. 691-700.

Manning, C. D., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT press.

McAuley, J., and Leskovec, J. 2013. "Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text," *Proceedings of the 7th ACM conference on Recommender systems*: ACM, pp. 165-172.

McCallum, A., and Nigam, K. 1998. "A Comparison of Event Models for Naive Bayes Text Classification," *AAAI-98 workshop on learning for text categorization*: Citeseer, pp. 41-48.

Mudambi, S. M., and Schuff, D. 2010. "What Makes a Helpful Review? A Study of Customer Reviews on Amazon. Com," *MIS quarterly* (34:1), pp. 185-200.

Nelson, P. 1970. "Information and Consumer Behavior," *The Journal of Political Economy*), pp. 311-329.

Otterbacher, J. 2009. "'Helpfulness' in Online Communities: A Measure of Message Quality," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: ACM, pp. 955-964.

Pang, B., and Lee, L. 2008. "Opinion Mining and Sentiment Analysis," *Foundations and trends in information retrieval* (2:1-2), pp. 1-135.

Provost, F., and Fawcett, T. 2001. "Robust Classification for Imprecise Environments," *Machine learning* (42:3), pp. 203-231.

Resnick, P., Zeckhauser, R., Swanson, J., and Lockwood, K. 2006. "The Value of Reputation on Ebay: A Controlled Experiment," *Experimental economics* (9:2), pp. 79-101.

Standifird, S. S. 2001. "Reputation and E-Commerce: Ebay Auctions and the Asymmetrical Impact of Positive and Negative Ratings," *Journal of Management* (27:3), pp. 279-295.

Tsaparas, P., Ntoulas, A., and Terzi, E. 2011. "Selecting a Comprehensive Set of Reviews," *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM, pp. 168-176.

Wu, P. F., Van Der Heijden, H., and Korfiatis, N. 2011. "The Influences of Negativity and Review Quality on the Helpfulness of Online Reviews," *International conference on information systems*.

Wyse, A. E., Keesler, V., and Schneider, B. 2008. "Assessing the Effects of Small School Size on Mathematics Achievement: A Propensity Score-Matching Approach," *The Teachers College Record* (110:9), pp. 1879-1900.

Ye, Y., and Kaskutas, L. A. 2009. "Using Propensity Scores to Adjust for Selection Bias When Assessing the Effectiveness of Alcoholics Anonymous in Observational Studies," *Drug and Alcohol Dependence* (104:1), pp. 56-64.

Yin, D., Bond, S., and Zhang, H. 2014. "Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews," *Mis Quarterly* (38:2), pp. 539-560.

Zhang, K. Z., Lee, M. K., and Zhao, S. J. 2010. "Understanding the Informational Social Influence of Online Review Platforms," *ICIS*, p. 71.

Zhou, W., and Duan, W. 2010. "Online User Reviews and Professional Reviews: A Bayesian Approach to Model Mediation and Moderation Effects," *ICIS*, p. 256.