Canary in the e-Commerce Coal Mine: Detecting and Predicting Poor Experiences Using Buyer-to-Seller Messages

DIMITRIY V. MASTEROV, eBay Research Labs
UWE F. MAYER, eBay Data Labs
STEVEN TADELIS, UC Berkeley Haas School of Business and NBER

Reputation and feedback systems in online marketplaces are often biased, making it difficult to ascertain the quality of sellers. We use post-transaction, buyer-to-seller message traffic to detect signals of unsatisfactory transactions on eBay. We posit that a message sent after the item was paid for serves as a reliable indicator that the buyer may be unhappy with that purchase, particularly when the message included words associated with a negative experience. The fraction of a seller's message traffic that was negative predicts whether a buyer who transacts with this seller will stop purchasing on eBay, implying that platforms can use these messages as an additional signal of seller quality.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Economics

General Terms: e-commerce, reputation systems, trust

Additional Key Words and Phrases: e-commerce; reputation systems; trust

1. INTRODUCTION

Online marketplaces such as eBay, Taobao, and Airbnb, to name a few, encompass an increasing share of economic activity and are rapidly growing worldwide. The anonymity of traders on platform markets raises concerns about asymmetric information. As a response to this potential hazard, e-commerce marketplaces use some sort of decentralized "reputation" or "feedback" mechanisms to alleviate these concerns and help buyers select trustworthy sellers (see, e.g., [Resnick et al. 2000; Dellarocas 2003] and [Cabral and Hortaçsu 2010]). In parallel, online sites such as Yelp create online feedback mechanisms that apply to brick-and-mortar shops, and these reputation scores have been shown to influence the performance of these establishments (see [Luca 2014]). It has been shown, however, that user-generated feedback mechanisms are often biased, and can be prone to influence by sellers (see, e.g., [Bolton et al. 2012; Dellarocas 2000; Dellarocas and Wood 2008; Fradkin et al. 2014; Mayzlin et al. 2014] and [Nosko and Tadelis 2015] for recent studies.) As a consequence, the information contained in online reputation and feedback mechanisms will not accurately convey the quality of sellers, which in turn will result in some buyers who rely on feedback to engage in transactions that do not meet their expectations.

In this paper we argue that a marketplace can use data that is generated in the natural course of activities that occur between buyers and sellers to create independent measures of seller quality, thus helping the platform distinguish between sellers who provide a better buyer experience, and those who are more likely to cause an unpleasant buyer experiences. In particular, we propose that transactions that go poorly

Author's addresses: D. Masterov, eBay Research Labs; U. Mayer, eBay Data Labs; S. Tadelis, UC Berkley Haas School of Business and NBER.

EC'15, June 15–19, 2015, Portland, OR, USA. Copyright © 2015 ACM 978-1-4503-3410-5/15/06 ...$15.00.
http://dx.doi.org/10.1145/2764468.2764499

are more likely to prompt a message from the buyer to the seller *after* the transaction was completed. Furthermore, messages with content that clearly identifies a poor experience signify that the buyer was unhappy with the transaction, suggesting that the seller provided a sub-standard level of service that leaves the buyer unhappy.

To prove our point we begin our analysis by taking a simple econometric model to the contents of post-transaction buyer-to-seller (B2S) message traffic in order to detect reported poor experiences (PEs) on eBay. Our analysis confirms that the existence of *any* post-transaction (post payment) messages serves as a reliable indicator that the buyer may be unhappy with that purchase, particularly when the message was a *negative* one. In particular, both simple correlations and regression analyses show that post-transaction B2S messages are associated with PEs reported on eBay's platform, and more so when the content of the message is negative.

Using this finding, we proceed to create a history-based measure of "seller quality" that is constructed from the messages sent to the seller after each transaction was completed. More precisely, for each seller we construct a message-based quality measure of the *fraction* of transactions the seller completed that were followed by a negative B2S message. We then use an econometric test to confirm that sellers who are of lower quality based on this measure are more likely to cause a buyer to leave the site after a transaction with that seller. This is a "revealed preference" approach similar to that in [Nosko and Tadelis 2015], which is based on the notion that a buyer who suffered a poor experience will choose to "exit" and vote with his feet so as not to return to eBay. Interestingly, the fraction of a seller's message traffic that was negative served as a less robust predictor that the seller would continue to create PEs in subsequent interactions with buyers.

This is an important observation because it sheds light on the problems, primarily that of bias, of user-generated feedback. The premise of much of the work on online reputation systems is that user-generated feedback plays an important role in facilitating trade. However, this study, similar to the observations in [Nosko and Tadelis 2015], shows that it is possible for e-commerce marketplaces to use naturally occurring data, that is not solicited as user-generated feedback, to better assess the quality of transactions and their impact on buyer retention.

It is important to note that the use of B2S message traffic alone is unlikely to be the optimal unique input to determine the quality of transactions. Similar to the approach in [Nosko and Tadelis 2015], we view our exercise here as a "proof of concept" and show how even a simple use of naturally occurring data, such as B2S messages, can improve a marketplace's ability to predict the quality of transactions and identify sellers that cause buyers to leave the marketplace platform. It surely will be the case that different platforms will likely have different sources of naturally occurring data that can be used to estimate the quality of transactions. Hence, the form of the optimal estimator in different platform markets is a question of statistical fit and engineering, informed by economic theories of buyer and seller behavior.

Being the first successful online marketplace, eBay's use of its reputation system is credited for its success, and it established a model for many marketplaces that followed (see [Dellarocas 2001]). As such, we are confident that our findings, and the approach we advocate, are relevant outside the eBay context and apply to many online marketplaces and two-sided platforms. Many reputation and feedback systems allow transacting parties to use the threat of a negative review to reach compromises. Insofar as this process is successful, and the review never materializes, this mechanism obscures the evidence that something went awry from the platform and this information is not used to alter how these sellers are presented to consumers, either by re-ranking them in the search results or by removing them from the platform entirely. The messaging data is

just one example of how internal information could be used to improve the reputation system, while still minimizing load on the platform from dispute arbitration.

2. DATA DESCRIPTION

We use a 20% sample of U.S. buyers who had any transactions in June of 2011 on the US eBay site. If a buyer had more than one transaction, we used a case-control sampling scheme where we picked a random transaction from the group of transactions that were designated as causing a poor experience (PE) for the buyer. During this period, PEs are detected whenever one of the following five buyer-generated events happen: (1) a buyer claims that the item was not received (INR); (2) a buyer claims that the item was significantly not as described (SNAD); (3) a buyer attempts a PayPal "chargeback" though their linked funding source (CB); (4) a buyer leaves negative or neutral feedback (N/N FB); (5) a buyer leaves low detailed seller ratings (DSRs) of item condition, communication, ship time or cost. If none of a buyer's transactions included a PE, we sampled a random transaction.[1] Then for each remaining buyer-seller-item triad, we observed whether the buyer sent the seller one or more messages post-transaction. Any messages sent before the transaction was completed—such as a question about a listing—were not included in the tally. If this triad engaged in more than one transaction in that month, we consider only the messages from the first one.

In order to classify the B2S messages, a regular expression search with a standard list of negative words was used to classify messages as negative or neutral. Negative messages include terms such as "annoyed," "dissatisfied," "damaged," or "negative feedback." If none of these terms appeared, the message was considered neutral. Using this classification we then grouped the transactions into 3 distinct types:

(1) No post-transaction messages from buyer to seller
(2) One or more negative messages
(3) One or more neutral messages with no negative messages

The spineplot in Figure 1 describes the distribution of transactions with the different B2S message classifications, and their association with PEs. The x-axis of the spineplot in Figure 1 shows that approximately 85% of transactions fall into the benign first category of no post-transaction B2S messages.

Buyers sent at least one message in the remaining 15% of all transactions, with the mixture being evenly split between negative and neutral B2S messages. The top of the y-axis shows the poor experience rate for each message type in blue. When no messages are exchanged, only 4% of buyers report a PE. Whenever a *neutral* message is sent, the PE rate jumps to 13%. If the content of the message was instead *negative*, over a third of buyers proclaim their dissatisfaction.

B2S message content is also positively correlated with refunds and the various subtypes of PEs. On one hand, the first aspect is mechanical, since synonyms like "refund" and "money back" were on the list of negative regular expressions. On the other hand, it also suggests that even though the regular expressions approach is primitive, it is allowing us to separate experiences that are qualitatively different. The spineplot in Figure 2 shows that if we include refunds, there's almost a 50% chance that a negative message indicates a future problem.[2]

[1] This sampling scheme ensures that anyone who experienced any PEs is tagged as such for the purpose of the analysis. It is also conservative, insofar as any misclassification of the binary outcome will attenuate the effects that we seek to estimate. This ensures that our estimate is a lower bound on the true effect.

[2] For this figure, we only counted the most severe problem for each transaction since a PE can be triggered by a combination of aforementioned causes. For example, if a customer left a low communication detailed seller rating and also reported a SNAD, only the latter would count for the purpose of this graph. This graph

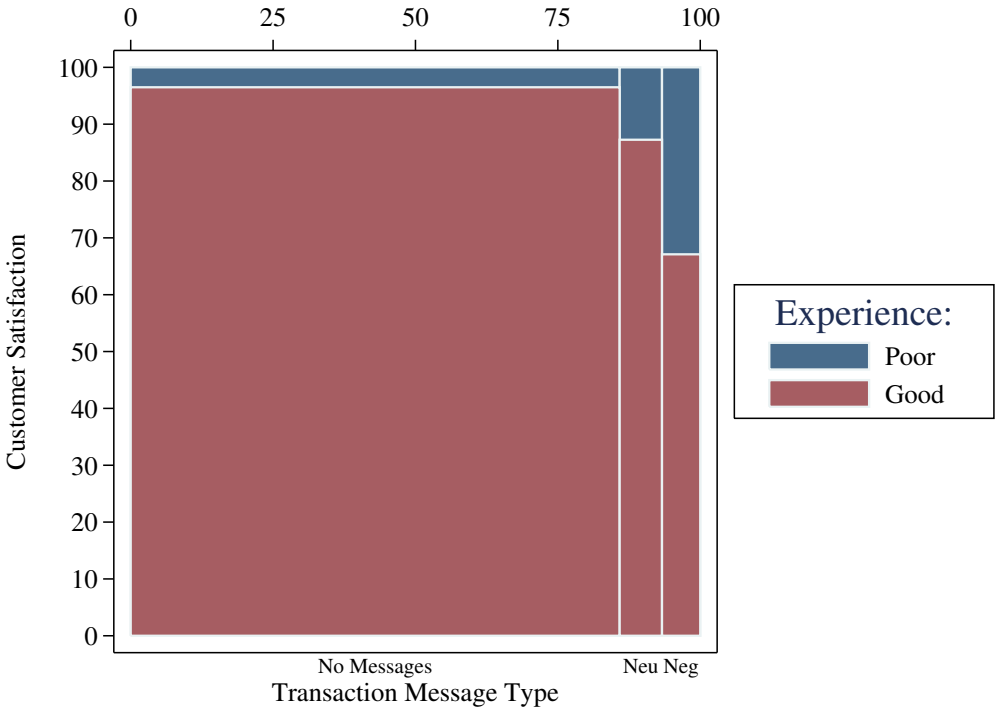


Fig. 1. Distribution of PEs by Message Type

3. CONSTRUCTING A B2S QUALITY SCORE

After establishing a positive correlation between B2S messages and PEs, especially for negative B2S messages, we proceed to use these messages to construct a new measure of seller quality.[3] The premise is that sellers who generate a higher frequency of negative B2S messages are worse sellers. Hence, we constructed a measure of seller quality based on the *fraction* of all the seller's transactions that had one or more negative B2S messages, and we call this measure "B2S Quality Score."

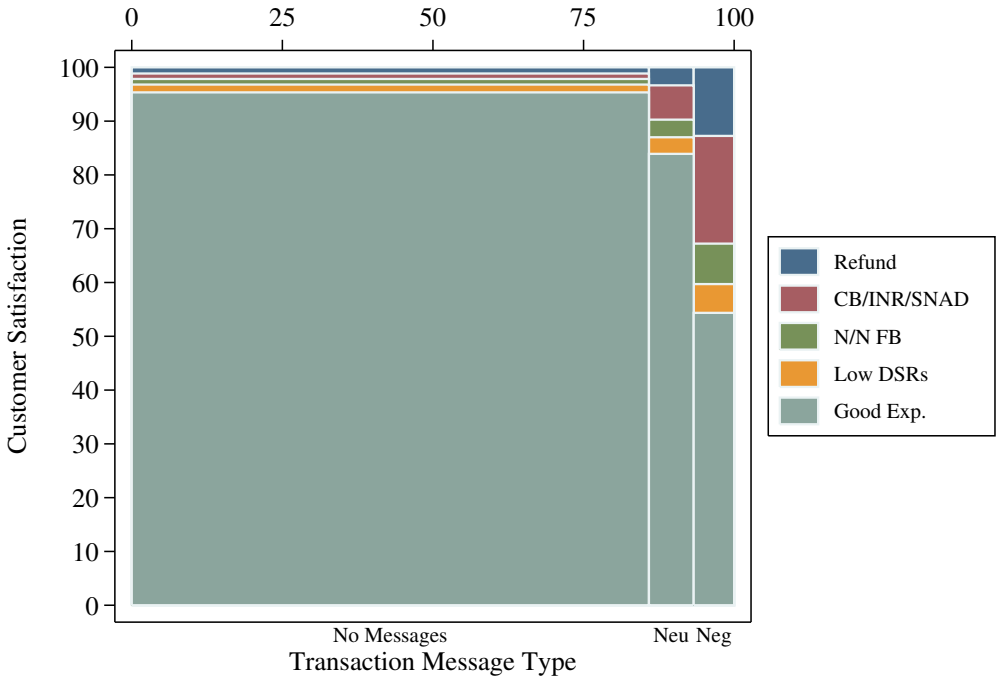
As an illustration, imagine that seller A and seller B both sold 100 items. Imagine further that of seller A's 100 transactions, five transactions had least one negative B2S message, while for seller B there were eight such transactions. The B2S quality score of seller A is then 0.05 while the B2S score of seller B is 0.08. The premise is that seller B, who has a higher B2S quality score, is a worse seller than seller A.

The relationship between this ratio, which is calculated using aggregated negative messages from past sales, and the likelihood that a current transaction will result in a PE, is monotonically increasing over most of the relevant range as shown in Figure 3. Only relatively recent transactions between June 2010 and June 2011 were used

---

also shows that it is possible to obtain a refund through a non-message channel, perhaps through contact information that was shipped with the item or because eBay sometimes exposes the e-mail addresses of the seller on the site.

[3]As we explain in the introduction, more complex measures can and should be constructed, but as the analysis below demonstrates, even the simple measure we construct contains valuable information that significantly augments other measures of seller quality.



Severe outcomes trump lesser ones.

Fig. 2. PEs and Refund Rates by Message Type

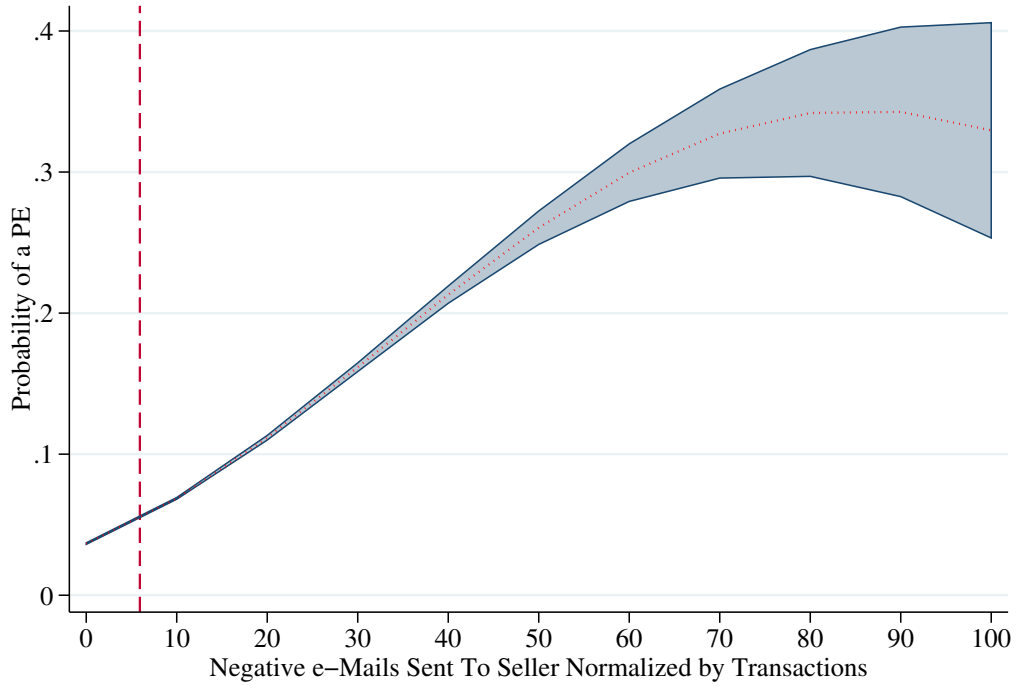
to calculate this metric. The marginal effect of a 10 unit increase in this metric is a 3 percentage point higher likelihood of a PE. Figure 3 suggests that high values of the B2S quality score can be used to flag sellers that cause PEs and as a result may disengage buyers. This idea is carefully explored using regression analyses in the next two sections.

4. REGRESSION RESULTS USING B2S MESSAGE CONTENT

In this section, we present regression results that go beyond the simple spineplots shown above to take into account various characteristics of the item, the buyer, and the seller. The control variables that were included in all the specifications are listed in Table I. It is important to note that because B2S messages, as well as the quality measure we construct are not observable to buyers, then buyers cannot select on these measures and we can interpret the results of our empirical analysis as causal.

The approach we take uses a probit index function model to represent the probability that a PE would be reported. The model assumes that there exists a *latent* dissatisfaction variable, y^*_{ijt}, that measures buyer i's dissatisfaction with a transaction t bought from seller j that is a linear-in-parameters function of several observable variables, the type of B2S message (if one was sent for this transaction), and an unobserved noise variable. The latent dissatisfaction variable y^*_{ijt} can be written as follows,

$$y^*_{ijt} = \beta \cdot \bar{b}_{it} + \gamma \cdot \bar{s}_{jt} + \delta \cdot \bar{d}_t + \lambda p_i + \rho n_i + \varepsilon_i,$$



Sample mean shown by vertical dashed line. Graph excludes sellers with fewer than 5 transactions in the previous year.

Fig. 3. Probability of PE by B2S Quality Score

where $\bar{b}_{it}$ is a vector of buyer characteristics (e.g., how experienced the buyer is, has the buyer also sold before, what the buyer's income is, etc.), $\bar{s}_{jt}$ is a vector of seller characteristics (e.g., whether the seller is new, whether they have a good track record, etc.), and $\bar{d}_t$ is a vector of transaction characteristics (e.g., whether the item is new or used, whether the selling format is an auction, etc.). The full list of control variables that we use in our regressions is listed in Table I. In addition to buyer, seller and transaction characteristics, p_i and n_i are binary indicators that at least one neutral or negative post-transaction message was sent, respectively, and ε_i is the error term that captures the influence of unobserved variables or measurement error of the PE reporting process. Interpretations of ε_i can be the cost of reporting a PE following a poor experience or the effort of contacting customer service, which varies across the population of consumers (some buyers may not take the time to report a PE and just move on).

Note that for seller characteristics we do not use the seller's feedback scores or percent positive, but instead use a variable called Effective Percent Positive, or EPP. [Nosko and Tadelis 2015] use eBay data to create the EPP score—a measure of seller quality unobserved by the buyer—and show that sellers with lower EPP scores are more likely to cause buyer dissatisfaction.[4] As [Nosko and Tadelis 2015] show, the

[4]A seller's EPP is defined as the number of positive feedback transactions divided by total transactions, thus holding "silence" against a seller's track record. [Nosko and Tadelis 2015] show that this measure is not only negatively correlated with PEs, but it predicts the likelihood of a buyer to return to eBay after a transaction, where sellers with lower EPP scores cause buyers to leave the platform with a higher likelihood.

Table I. Explanatory Variables

Seller	*Buyer*	*Item / Transaction*
New Seller	New Buyer	Auction Format
New Seller × New Buyer	Time Since First Bid/Purchase	Used Condition
Log of Previous Transactions + 1	Buyer Also a Seller	Meta Category
Effective Percent Positive $= \frac{\text{Transactions With Positive FB}}{\text{Number of Transactions}}$	Income Group	Day of the Week
	Gender, Age, Educational Level	
	Num. of Adults in HH	
	Num. of Children in HH	
	Married vs Single HH	
	Occupation	
	Home Owner vs Renter	
	Dwelling Type	

EPP score contains a lot more information than either of the commonly used feedback and reputation scores, which is why we omit them form our analysis. Including them has almost no impact on the magnitude of our results.

We cannot, however, directly observe buyer dissatisfaction. Instead, all we see is whether a PE was reported by the buyer. Our approach assumes that a PE is reported by a buyer when the level of the latent dissatisfaction variable, y^*_{ijt}, exceeds a certain threshold that makes the customer willing to report a PE:

$$y_{ijt} = \begin{cases} 1 & \text{if } y^*_{ijt} > 0 \\ 0 & \text{if } y^*_{ijt} \leq 0 \end{cases} .$$

Note that a threshold of zero is an innocuous normalization. In what follows we will drop the individual subscripts ijt to avoid cluttered notation. Also, to simplify further, we slightly abuse notation and collapse the set of covariates $(\beta \cdot \bar{b}_{it} + \gamma \cdot \bar{s}_{jt} + \delta \cdot \bar{d}_t)$ and replace them with the simplified expression $x'\beta$.

We can model this binary outcome as

$$\begin{aligned} \Pr[y = 1 \mid x, p, n] &= \Pr[y^* > 0] \\ &= \Pr[x'\beta + \lambda \cdot p + \rho \cdot n + \varepsilon > 0] \\ &= \Pr[-(x'\beta + \lambda \cdot p + \rho \cdot n) < \varepsilon] \\ &= F(x'\beta + \lambda \cdot p + \rho \cdot n) \end{aligned}$$

where F is the cumulative density function of $-\varepsilon$, which is also the CDF of ε as long as ε is symmetric about zero. If we assume that $\varepsilon \sim N\left(0, \sigma^2\right)$, then the conditional probability of a PE becomes

$$\begin{aligned} \Pr[y = 1 \mid x, p, n] &= \Phi\left(\frac{x'\beta + \lambda \cdot p + \rho \cdot n}{\sigma} > -\frac{\varepsilon}{\sigma}\right) \\ &= \Phi\left(x'\frac{\beta}{\sigma} + \frac{\lambda}{\sigma} \cdot p + \frac{\rho}{\sigma} \cdot n\right), \end{aligned}$$

where Φ is the standard normal CDF. We have normalized the variance of the error to be one. It follows that

$$\Pr[y = 0 \mid x, p, n] = 1 - \Phi\left(x'\frac{\beta}{\sigma} + \frac{\lambda}{\sigma} \cdot p + \frac{\rho}{\sigma} \cdot n\right),$$

Another boon of this measure is that unlike the percentage positive reported on the site, there is considerable variation across sellers.

so that we can numerically estimate the parameters (up to scale) using maximum likelihood estimation with

$$\mathcal{L} = \prod_{i=1}^{N} \left(1 - \Phi\left(x'\frac{\beta}{\sigma} + \frac{\lambda}{\sigma}\cdot p + \frac{\rho}{\sigma}\cdot n\right)\right)^{1-y_i} \cdot \Phi\left(x'\frac{\beta}{\sigma} + \frac{\lambda}{\sigma}\cdot p + \frac{\rho}{\sigma}\cdot n\right)^{y_i},$$

after taking the natural log of the likelihood. This function is globally concave, so the estimation is fairly straightforward.[5]

We are not particularly interested in the effect of p and n on the latent dissatisfaction y^*, which is complicated by the fact that dissatisfaction does not have a well-defined unit of measurement. Instead, we will be concerned with the marginal effect of p and n on the probability of reporting a PE, $\Pr[y = 1 \mid x, p, n]$, relative to having no B2S messages. Since these variables are binary, we will be providing the following average marginal effects (AME) and their confidence intervals, calculated by taking finite differences:

$$\begin{aligned}
AME_p &= \frac{1}{N}\sum_{i=1}^{N}\left[\Phi\left(x_i'\beta + \lambda\cdot 1\right) - \Phi\left(x_i'\beta + \lambda\cdot 0\right)\right] \\
AME_n &= \frac{1}{N}\sum_{i=1}^{N}\left[\Phi\left(x_i'\beta + \rho\cdot 1\right) - \Phi\left(x_i'\beta + \rho\cdot 0\right)\right] \\
Baseline &= \frac{1}{N}\sum_{i=1}^{N}\Phi\left(x_i'\beta\right)
\end{aligned}$$

[5] It is also possible to derive the probit specification from a simple choice model. Suppose we have a customer who is deciding whether to report a PE or not, possibly dishonestly. The utilities he receives from the two alternatives are

$$\begin{aligned}
U_1 &= x'\alpha_1 + u_1 \\
U_2 &= x'\alpha_2 + u_2,
\end{aligned}$$

where we assume that both the error terms u_j are normally distributed and independent, x is a vector of explanatory variable, and α_k are unobserved utility parameters. The net utility from reporting is

$$U_1 - U_2 = x'\left(\alpha_1 - \alpha_2\right) + \left(u_1 - u_2\right),$$

so the agent will chose to report as long as $(U_1 - U_2) > 0$. The probability of reporting is then

$$\begin{aligned}
\Pr\left[U_1 > U_2\right] &= \Pr\left[x'\left(\alpha_1 - \alpha_2\right) > -\left(u_1 - u_2\right)\right] \\
&= \Pr\left[x'\left(\alpha_1 - \alpha_2\right) < \left(u_1 - u_2\right)\right].
\end{aligned}$$

The second step above follows from the symmetry of the normal distribution and the fact that a difference of two normal variables is itself normally distributed. Dividing by the standard deviation of the error difference yields

$$\begin{aligned}
\Pr\left[U_1 > U_2\right] &= \Pr\left[x'\frac{\left(\alpha_1 - \alpha_2\right)}{\sigma} < \frac{\left(u_1 - u_2\right)}{\sigma}\right] \\
&= \Phi\left(x'\frac{\left(\alpha_1 - \alpha_2\right)}{\sigma}\right) \\
&= \Phi\left(x'\frac{\beta}{\sigma}\right).
\end{aligned}$$

The last equation is the baseline PE rate when no messages are exchanged, and is intended to give a sense of scale for the first two. The standard errors used in calculating the confidence intervals were calculated using the delta method since Φ is a nonlinear function.

4.1. Baseline Average Marginal Effects

The average marginal effects are shown in Table II. All the continuous variables have been re-scaled to be mean zero, so that the baseline is the PE rate for the person with the average or modal covariates and no B2S messages.

Table II. Average Marginal Effects of B2S Messages on PE Rate

	AME & 95% CI
1+ Neutral Messages	0.075^{***}
	[0.072,0.077]
1+ Negative Messages	0.260^{***}
	[0.256,0.264]
Baseline (No Messages)	$.027^{***}$
	[.026,.027]
Obs	773,283
McFadden's Adjusted R^2	19.1

Model includes all covariates from Table I

This baseline rate is 2.7%. If we see a neutral message, the PE rate jumps up by 7.5 percentage points. This means a tenth of these transactions are "bad." If the message had negative content, the increase is 26 percentage points, so that almost a third of such transactions would lead to a reported PE. All the $AMEs$ are precisely estimated, and the area under the ROC curve is 0.8. Note that these effects are holding the EPP constant, which means that the message content carries additional signal beyond what is in the EPP measure.

The fact that neutral messages predict unsatisfactory transactions may seem curious, but should not be too much of a surprise. After all, it is reasonable to assume that a successful transaction should not trigger any need for a buyer to contact the seller. One possible explanation is that the regular expression search is a maladroit classification method. It is likely that a more sophisticated natural language processing (NLP) algorithm would yield better results than frugal content heuristics, and identify more messages as having some kind of negative content.

As we mention earlier in the paper, we refrain from a more sophisticated NLP analysis as the goal of this paper is to demonstrate the gains from using B2S messages, and not necessarily constructing the most efficient one. The optimal signal of seller quality would be found using a host of machine-learning methods and would, of course, be platform-specific.

4.2. Additional Average Marginal Effects

The effects of message content shown above are much larger than those for any other variables in the model. Some of the substantial ones are shown in Table III.

For example, a PE is 1.1 percentage points less likely with a seller who is one standard deviation higher in EPP, the measure of seller quality identified and used in [Nosko and Tadelis 2015]. Similarly, a buyer who also sells on eBay is 1.6 percentage points more likely to report a PE. An effect of comparable magnitude is found for the auction format. Interestingly, used items are only marginally more likely to create dissatisfaction at 0.6 percentage points, though this is conditional on meta category. Finally, while these estimates are small in absolute terms, they are large in relative terms given how low the baseline PE rate truly is.

Table III. The Average Marginal Effect of Selected Covariates

Variables	AME on $\Pr(PE)$
One Std. Dev. Increase in EPP	-0.0111^{***}
Buyer Also a Seller	0.0157^{***}
Auction	0.0164^{***}
Used Item	0.0064^{***}
Obs	773,283

Model includes all covariates from Table I

5. REGRESSION ANALYSIS USING B2S QUALITY SCORE

This section discusses the average marginal effects from the regression using the B2S Quality Score, which is the fraction of all historical transactions that had one or more negative B2S messages in the previous year.

5.1. PE Regressions

Recall that in our model, the probability of a PE is

$$\Pr\left[y_i = 1 \mid x_i, s_i\right] = \Phi\left(x_i'\beta + \omega \cdot s_i + \chi \cdot s_i^2\right),$$

where s is the B2S quality score. A quadratic function of the score was used to capture the non-linearity in the effect we saw in the graph. We are going to present the average marginal effect of s, defined as

$$\begin{aligned} AME_s &= \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \Pr\left[y_i = 1 \mid x_i, s_i\right]}{\partial s} \\ &= \frac{1}{N}\sum_{i=1}^{N} \phi\left(x_i'\beta + \omega \cdot s_i + \chi \cdot s_i^2\right) \cdot \left(\omega + 2 \cdot \chi \cdot s_i\right). \end{aligned}$$

This is a function of s, so we will evaluate it at several parts of the normalized score distribution. Table IV shows the $AMEs$.

Table IV. Average Marginal Effect of B2S Quality Score on PE

	AME & 95% CI
3 SD Below	0.0082^{***}
	[0.0081,0.0083]
2 SD Below	0.0111^{***}
	[0.0108,0.0113]
1 SD Below	0.0137^{***}
	[0.0132,0.0142]
At Mean	0.0157^{***}
	[0.0149,0.0164]
1 SD Above	0.0167^{***}
	[0.0158,0.0175]
2 SD Above	0.0165^{***}
	[0.0155,0.0174]
3 SD Above	0.0151^{***}
	[0.0141,0.0161]
Baseline	.053
	[.0525,.0535]
Obs	773,283

Model includes all covariates from Table I

For example, consider a buyer who has purchased from a seller who has the average negative message score. If that seller was one standard deviation above the mean, the

likelihood of a PE would increase by 1.57 percentage points above the baseline. The effects are weaker near the bottom and of comparable magnitude near the top, with some of the dip we saw in the simple graph.

As expected, these effects are smaller in magnitude than with the actual transaction message because this is an aggregate measure. However, this historical data is available *before* the transaction occurs. This means that the marketplace platform can use an algorithm to calculate the B2S quality score and rank sellers according to their score ahead of time for search presentation.[6] We discuss this further in the conclusions.

5.2. Buyer Exit Regressions

The analysis we conducted so far considers a measures of buyer satisfaction, namely the PE identifier, as they relate to our B2S quality score. As much as these relationship are suggestive of a seller's quality, they do not capture the impact of a seller's quality on whether or not buyers choose to respond to bad experiences beyond reporting a problem. E-commerce platforms will suffer from PEs and from lower quality sellers only to the extent that these will impose a loss of business, and as a result, a platform experience that is plagued by asymmetric information is a form of market failure. An important question is whether our measures of seller quality can actually predict platform abandonment by buyers.

Note that buyers *do not* observe the B2S quality score of sellers, and hence, cannot select on it. As a consequence, we can consider a regression of whether or not a buyer chooses to leave the site on a host of control variables as well as our B2S quality score. It is precisely because the B2S quality is not observed by the buyers that this regression will result in a causal inference that is well identified. It is also worth noting that using whether or not a buyer remains active is a "revealed preference" approach that identifies buyer sentiment not by their user-generated content, but whether they vote with their feet, as in [Nosko and Tadelis 2015]. The other empirical papers that use data from eBay and similar marketplaces are constrained to see the effect of reputation scores, which suffer from biases, on variable such as prices and quantities, and these do not capture the negative impacts of bad transactions on the retention of buyers.

Table V below reports marginal effects on the likelihood that a seller leaves the platform after a transaction from changing either EPP or the B2S quality score by one standard deviation, while keeping the other constant. The left-side (dependent) variable of the regression is an indicator variable of whether or not the buyer decided to return to eBay *after* the transaction, and the right-side (independent) variables include EPP and B2S quality score, as well as other buyer, seller and item controls. A buyer who left the site after a transaction is defined as making zero purchases within one year of the target June 2011 transaction. If at least one transaction was made, the buyer is considered stable. The baseline rate of leaving eBay is about 8%, though it is considerably higher for new customers (see [Nosko and Tadelis 2015]).

Both marginal effects are in the expected direction. Interacting with a seller who is one standard deviation above the mean of the EPP distribution appears to reduce the likelihood that a buyer leaves the platform by 0.8 percentage points, which is approximately 10%. This is reasonable since [Nosko and Tadelis 2015] show that EPP is a valid measure of seller quality. The effect is somewhat stronger in the left tail of the distribution and a bit weaker in the right, suggesting diminishing returns to seller quality. This marginal effect implicitly holds the B2S Ratio constant.

[6]Other potential extensions that we do not consider here are using a dollar-weighted metric rather than treating all negative messages equally, and weighting more recent transactions more heavily than distant ones in constructing the B2S score. As mentioned earlier, we are not trying to construct the optimal signal but rather show that our approach is feasible and adds valuable information.

Table V. Average Marginal Effect On Exit From a 1 Std. Dev. Increase In B2S Score or EPP

	B2S Score	EPP
3 SD Below	0.009***	0.006***
	[0.008,0.009]	[0.003,0.008]
2 SD Below	0.009***	0.001
	[0.008,0.010]	[-0.001,0.003]
1 SD Below	0.008***	-0.004***
	[0.007,0.009]	[-0.005,-0.003]
At Mean	0.008***	-0.008***
	[0.007,0.009]	[-0.009,-0.007]
1 SD Above	0.007***	-0.011***
	[0.006,0.008]	[-0.013,-0.010]
2 SD Above	0.006***	-0.013***
	[0.005,0.007]	[-0.015,-0.012]
3 SD Above	0.004***	-0.013***
	[0.004,0.005]	[-0.015,-0.012]
Obs	773,283	773,283

Model includes all covariates from Table I

A higher B2S Quality Score corresponds to a *lower* quality seller since it indicates more negative B2S message traffic. An increase of one standard deviation at the mean of this measure increases the likelihood that a buyer leaves the platform by 0.8 percentage points. This is the same magnitude as with EPP. The same type of concave relationship between seller quality and churn is evident. What is interesting is that EPP and our B2S quality score complement each other as they are not highly correlated. This establishes that platforms may have access to many internally constructed measures of seller quality that can be used to improve the platform and increase buyer retention.

6. CONCLUSION

If a buyer has a poor experience on an e-commerce platform, this should cause the buyer to be less likely to continue engaging and purchasing from the site. This hurts both consumer surplus and platform profits, as these problems are extremely detrimental to customer lifetime value, particularly for new users. As a result, platform marketplaces have tried to use a variety of feedback and reputation mechanisms to help buyers find reliable sellers to transact with. Many have argues that the feedback will provide valuable information for future buyers (e.g., [Gregg and Scott 2008]) as it can reduce informational asymmetries that are caused by adverse selection and moral hazard (see [Klein et al. 2013]).

We have argued and shown that platforms can use other measures of seller quality that can be constructed using naturally occurring data from buyer and seller engagement, focusing attention on B2S messages. The presence of these messages allows us to detect problems early and to take steps to potentially offset the damage that poor quality sellers can cause. Platforms can then use these measures of seller quality to demote or promote sellers that are deemed to be of low or high quality. More broadly, the contents of message traffic can be used to bolster the reputation and feedback mechanism in valuable ways.

Our results are based on data from eBay, one of the largest and arguably oldest e-commerce marketplaces. Other marketplaces such as Airbnb also allow for B2S messages, and it is likely that they too convey information that could augment the sometimes “too nice” ratings that are common in feedback mechanisms. Hence, our study offers new directions that online platform markets can follow in order to enhance buyer experience, and create longer lasting relationship and higher values for their consumers.

ACKNOWLEDGMENTS

The authors would like to thank Mark Boyd and Chris Nosko for early discussions that inspired us to explore this topic.

REFERENCES

Gary Bolton, Ben Greiner, and Axel Ockenfels. 2012. Engineering Trust: Reciprocity in the Production of Reputation Information. *Management Science* 59, 2 (2012), 265–285.

Luis Cabral and Ali Hortaçsu. 2010. The Dynamics of Seller Reputation: Evidence from eBay. *The Journal of Industrial Economics* 58, 1 (2010), 54–78.

Chrysanthos Dellarocas. 2000. Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior. In *Proceedings of the 2Nd ACM Conference on Electronic Commerce (EC '00)*. ACM, New York, NY, USA, 150–157. DOI:http://dx.doi.org/10.1145/352871.352889

Chrysanthos Dellarocas. 2001. Analyzing the Economic Efficiency of eBay-like Online Reputation Reporting Mechanisms. In *Proceedings of the 3rd ACM Conference on Electronic Commerce (EC '01)*. ACM, New York, NY, USA, 171–179. DOI:http://dx.doi.org/10.1145/501158.501177

Chrysanthos Dellarocas. 2003. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science* 49, 10 (2003), 1407–1424.

Chrysanthos Dellarocas and Charles Wood. 2008. The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias. *Management Science* 54, 3 (2008), 460–476.

Andrey Fradkin, Elena Grewal, David Holtz, and Matthew Pearson. 2014. Reporting Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on Airbnb. (2014). working paper.

Dawn G. Gregg and Judy E. Scott. 2008. A Typology of Complaints About eBay Sellers. *Commun. ACM* 51, 4 (April 2008), 69–74. DOI:http://dx.doi.org/10.1145/1330311.1330326

Tobias Klein, Christian Lambertz, and Konrad Stahl. 2013. Adverse Selection and Moral Hazard in Anonymous Markets. *CEPR Working Paper* DP9501 (2013).

Michael Luca. 2014. Reviews, Reputation, and Revenue: The Case of Yelp.com. *Working paper* (2014).

Dina Mayzlin, Yaniv Dover, and Judith Chevalier. 2014. Promotional Reviews: An Empirical Investigation of Online Review Manipulation. *American Economic Review* 104, 8 (2014), 2421–55.

Chris Nosko and Steven Tadelis. 2015. The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment. (2015). NBER working paper No. 20830.

Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation Systems. *Commun. ACM* 43, 12 (Dec. 2000), 45–48. DOI:http://dx.doi.org/10.1145/355112.355122