Asymmetric Information, Adverse Selection and Online Disclosure: The Case of eBay Motors

Gregory Lewis

Web Appendix

1 Theory Model

In the main text of the paper, I argued that there is a relationship between disclosure costs and market participation, because of the potential for adverse selection. The intuition is that with zero disclosure costs, everyone discloses and there is effectively no information asymmetry. On the other hand, with infinite disclosure costs, no-one discloses and there may be adverse selection. One might suspect that the case with positive costs is "in-between", and that adverse selection is increasing in disclosure costs. Here I provide a simple price-taking model that formalizes this intution.[1]

Let q denote the quality of a car. There are many risk-neutral buyers, and they all have willingness to pay for a car equal to its quality. Sellers value their car at $q - \alpha$, where $\alpha > 0$ and so the gains to trade are always positive. Sellers know the quality of their car, and can disclose it at cost c. Buyers do not know the car quality unless it is disclosed, and thus must form expectations of its quality conditional on non-disclosure. Buyers are also assumed to have no market power, and buy whenever the price is less than or equal to the (expected) quality.

Sellers differ along two dimensions: the quality of their car q and the cost of disclosure c. This defines their type. Their disclosure strategy is a mapping from their type to their decision to disclose. It is relatively straightforward to show that strategies must be monotone on both dimensions: (i) for fixed c, if (q_0, c) discloses then so will (q_1, c) for $q_1 > q_0$; (ii) for fixed q if (q, c_1) discloses, then so will (q, c_0) for $c_0 < c_1$. This implies threshold strategies. They may also choose whether to sell/list the car or not. They will list the car if buyers are willing to pay at least $q - \alpha$.[2]

[1] A more general version that explicitly accounts for the eBay auction format is available from the author on request.

[2] They could also try sell it at an unrealistic price; but since there is no uncertainty about what buyers are willing to pay in this model, an $\epsilon > 0$ participation cost would eliminate this possibility.

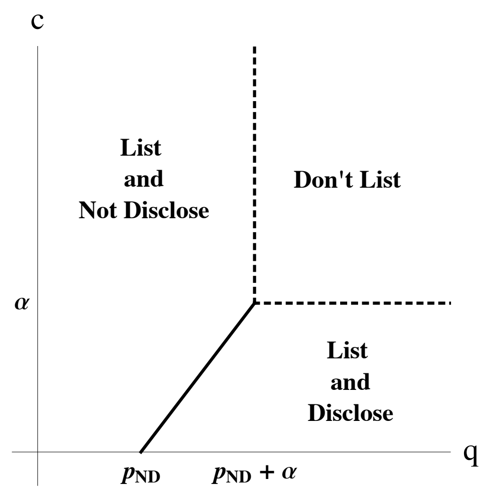

Figure 1: **Listing and Disclosure.** The figure shows how sellers of varying quality cars and with varying disclosure costs will behave.

To begin the analysis, fix an equilibrium. Those sellers who list their cars and disclose will receive a price $p = q$. Those sellers who list but do not disclose will all receive a common pooled price, p_{ND}. So for a seller to be indifferent between disclosing and not disclosing requires that:

$$q - p_{ND} = c \tag{1}$$

so that the marginal gain to disclosing equals its cost. Conditional on disclosing, the seller will sell for q, getting gains to trade α but paying the disclosure costs c. Indifference requires:

$$\alpha = c \tag{2}$$

Finally, if the seller plans not to disclose, then he is indifferent between listing and not listing when:

$$p_{ND} = q - \alpha \tag{3}$$

i.e. when the price he gets is equal to his private valuation.

It is really easy to use (1)—(3) to show how the different types behave, as illustrated in Figure 1. Equation (1) defines a straight line in type-space, with slope 1 and x-intercept p_{ND}. Types to the right of this line find it worthwhile to disclose. Equation (2) defines a horizontal line at α, along which types who would disclose are indifferent about listing. Similarly, equation (3) gives a vertical line at $p_{ND}+\alpha$ for non-disclosers.

Together, these split the type space into three regions: a "list and disclose" region populated by sellers with both high quality vehicles and low disclosure costs; a "list and not disclose" region, populated almost entirely by low quality vehicles; and a "don't list" region of sellers who have high quality vehicles, but also high disclosure costs. One should therefore expect that the cars sold on eBay are either of low quality, or can be proven to be of high quality at relatively low cost to the seller.

What happens when disclosure costs decrease for all types, as when a platform improves its disclosure technology? There are two main effects. First, disclosure becomes relatively more attractive for all types, and for any fixed original c, the threshold quality needed for disclosure falls. This implies more disclosure, and a lower price p'_{ND} for the the remaining lower-quality types who do not disclose. This results in positive selection: sellers with high quality cars who previously did not list them because disclosure was too costly for them will now select into the market (types on the horizontal dashed line). Taken in reverse, an increase in disclosure costs leads to adverse selection, as argued in the paper. Second, non-disclosing types on the vertical dashed line now strictly prefer not to list, because the new pooling price p'_{ND} is lower. This further worsens the quality of the pool of non-disclosing sellers. As disclosure becomes cheaper, the pool of non-disclosing sellers collapses to only those selling lemons.

2 Text Analysis Methodology

In the text analysis section of the paper, I focus on three nouns that may be present in the item description: "rust", "scratch" and "dent", and on qualifiers attached to these nouns. Here I describe the process used to choose these particular nouns and the associated qualifiers. Based on the methodology found in the content analysis literature[3], I went through the following steps:

1. **Obtain sample corpus:** Using a random sample of my data consisting of around 2000 auctions, I formed a text file consisting of the item descriptions from all these auctions.[4]

2. **Analyze word frequency:** I formed a table with the frequency of all words

[3] See for example, Kimberly Neuendorf, *The Content Analysis Guidebook* (2002).
[4] In linguistics, a large sample of text is called a corpus.

used in the item description.[5] From this table, I isolated nouns that were relevant for the value of the car, and were frequently used. The nouns eventually chosen were among the most frequently used (for example "rust" had a frequency of 600, compared to around 1500 for "car").

3. **Analyze concordance:** For each word chosen, I looked at how it was used in context, by examining a list of the phrases surrounding each word (a "concordance"). Based on this, I formed lists of qualifiers and negations that were typically used. For example, for negations, I used "no", "not", "never", "nothing", "free" (as in "rust-free"), "zero"; for adjectives, I used "small", "minor" etc. I also chose to discard nouns that upon closer examination were not relevant (e.g. "receipt" is typically used in sentences like "the successful buyer will be provided with a receipt", rather than "I have all the original parts receipts").

4. **Extract data:** Using a Python script, I searched for each of the three nouns, and where they were found, did a secondary search for any of the qualifiers within 50 characters of the noun. I coded up dummies for the presence and absence of these variables.

5. **Categorize instances:** Finally, I categorized each instance as either noun not present, noun present but positively qualified (e.g."minor rust problems"), noun present and unqualified, and noun present and negatively qualified (e.g. "major rust problems"). These categorical variables are those used in the regressions.

3 Software and Observables

In the last section of the paper, I examine how software upgrades and downgrades affect the number of photos posted; and subsequently use software as an instrument for photos in a hedonic regression. Here I examine whether the decision to switch software is correlated with a change in the observable characteristics of the cars sold pre-and-post switch. I define a seller to have "switched" anytime they change software; define them to have "upgraded" anytime they previously used the standard eBay software and now use a professional listing platform; and to have "downgraded" anytime they

[5] I used TextSTAT, a text analysis tool freely available on the internet.

go from the professional software to the standard software. Then I form variables for "pre-switch" and "post-switch" for the periods before and after the switch, and similarly define pre-and-post upgrade and downgrade. Where a seller switched multiple times (29% of sellers who switch), I define the pre-and-post variables based on the first switch.

In Table 1 below, I collapse the dataset to a mean for each seller under each regime, and compare the averages before and after. For example, for people who switch software at least once, I compute the average characteristics of the cars they sell before and after switching; and take this as a single datapoint. Table 1 gives the summary statistics for the dataset formed in this way.

Looking at the table, we see that there is little change in observable characteristics pre-and-post switches, or upgrades. The case of downgrades is more worrying, as we can see that following a downgrade the average mileage of cars listed goes up by over 5000 miles, or 7%. This could be because of endogenous software downgrades, or just due to noise. To get a handle on this, we want to formally test the null that the observable characteristics are drawn from the same distribution before and after. And we don't want to collapse the data to a seller-level, since some sellers have far more transactions with them than others and therefore their pre and post means are measured with more precision.

To avoid this problem, I follow a two step procedure. For any characteristic z_{ik}, I first compute a seller specific mean $\bar{z}_k$ and then for each transaction, a residual $r_{ik} = z_{ik} - \bar{z}_k$. This first step tries to control for seller heterogeneity, hopefully leaving us with a sample of draws r_{ik} from a mean zero distribution with common variance under the null. Then for any pre and post period I compute a T-statistic as:

$$t_k = \frac{\bar{r}_k^{post} - \bar{r}_k^{pre}}{S_{pool}\sqrt{\frac{1}{n_{pre}} + \frac{1}{n_{post}}}}$$

where $\bar{r}_k^{post}$ is the mean residual over all transactions in the post-period; $\bar{r}_k^{pre}$ is the mean residual over all transactions in the pre-period; S_{pool} is an unbiased estimate of the pooled standard deviation of the residuals; and n_{pre} and n_{post} are the number of observations in the pre and post periods. This gives us a T-statistic for each of the 5 listed characteristics, under each of the three different software changes (switch, upgrade, downgrade). The T-statistics are reported in Table 2 below:

Table 1: Observables under Software Changes

Variable	Pre-Switch	Post-Switch	Pre-Up	Post-Up	Pre-Down	Post-Down
Miles	69267	71325	77665	77124	69280	75592
	(65038)	(60444)	(70854)	(66138)	(54969)	(60179)
Age (in years)	10.11	10.25	12.02	11.86	11.25	11.96
	(10.08)	(10.20)	(11.19)	(11.42)	(10.75)	(10.41)
% Manual	26.41	26.53	23.60	28.63	30.00	25.19
	(37.10)	(35.23)	(35.88)	(36.68)	(36.02)	(33.46)
% "Reliable Car"	23.95	24.31	19.82	20.27	20.15	21.14
	(37.99)	(35.22)	(36.47)	(35.01)	(33.84)	(34.99)
% Trucks	29.71	28.56	31.65	30.48	25.24	24.61
	(41.43)	(37.97)	(42.68)	(39.63)	(35.88)	35.95)
N	327	327	169	169	124	124

Summary statistics of mean characteristics of cars sold by sellers under paired regimes: pre-and-post software switch, pre-and-post software upgrade, pre-and-post software downgrade. Each sample in turn only includes switchers, upgraders etc. Standard deviations of the mean characteristics are given in parentheses.

None of the T-statistics is significant at the 10% level, so we fail to reject the null of no endogenous software choices. Though none is individually significant, it could be that they jointly are, and so we use a Hotelling T-squared test to test for the equality of all mean residuals across the two groups. Again, we fail to reject the null at 10%. Overall then, there is little evidence that the observables vary with software choice.

Table 2: Testing for Changes in Observables

Variable	Switches	Upgrades	Downgrades
Miles	1.22	-0.94	0.41
Age (in years)	0.43	-0.87	0.44
% Manual	-0.06	1.25	-0.20
% "Reliable Car"	-1.16	0.69	-0.15
% Trucks	0.07	-0.65	-0.08
F-test statistic	0.71	0.70	0.08

The table reports t-statistics for a test of equality of the means of the seller-demeaned characteristics pre-and-post switching, pre-and-post upgrades and pre-and-post downgrades, respectively; as well as the F test statistic for the joint hypothesis that all pre-and-post means are equal (formed by normalizing the Hotelling T-squared). None of the statistics is significant at the 10% level.