Experimental Evaluation of an eBay-Style Self-Reporting Reputation Mechanism

Kay-Yut Chen, Tad Hogg

Hewlett-Packard Laboratories

Abstract. We experimentally studied the effects of a eBay-style self-reporting reputation mechanism in an double-sided exchange economy in which participants have the option of not fulfilling their contracts. We found that submitted reports quite accurately reflected their transactions and this mechanism maintaining a high contract fulfillment rate. The inaccurate reports, which were about 5% of the total, were heavily biased towards bad ratings when the transaction is successful. This is strong evidence that the inaccurate reports were not results of random errors, but derived from an underlying behavior effect. Our experimental design allowed identifying the effect of reputation mechanism on endogenous market behavior.

1 Introduction

Reputation has been an important aspect of commerce since the emergence of exchange economies [1]. Reputations can ensure promised actions are taken without the expense of external enforcement mechanisms or third party monitoring such as credit card companies. The Internet and subsequent development of e-commerce allow an increasing number of small players to engage in buying and selling. eBay is a prime example of how small businesses, particularly those serving niche markets, can overcome the previously forbidding marketing costs and reach customers with relatively low information costs. This trend leads further to transactions that take place entirely via the Internet when the product or service itself can be delivered on-line in addition to using the Internet to identify, negotiate and pay for the transaction.

However, this environment increases the importance of establishing trust in a market where everyone can choose to be anonymous, people may only participate in a few transactions, each transaction may be of relatively low value, and transactions readily cross jurisdictional boundaries raising the difficulty of legal contract enforcement. eBay approached this issue, with some success, with their feedback mechanism in which participants rate the performance of the other party in their transactions.

Establishing trust through repeated interactions has been studied in several contexts, particularly with the iterated Prisoner's Dilemma [2]. These give rise to strategies, such as tit-for-tat, to ensure cooperation. Another example is the experimental study of the "lemon" market in which reputation substantially affects behavior [3]. Unlike the Prisoner's Dilemma scenario, people in a market can choose not to do business with those deemed untrustworthy, or offer different terms based on perceived level of trust. Large companies can spread risk among many transactions (e.g., insurance) so have predictability arising from averaging over many individuals. On the other hand, small-scale transactions on the Internet lack this feature, perhaps leading risk-adverse people to avoid transactions that could benefit both parties. Such avoided transactions reduce market efficiency and hence decrease the potential economic gains from Internet's reduction in information and transaction costs. Thus an important question is to what extent reputation mechanisms can aid such markets. Analysis of eBay-like markets suggests that although a feedback mechanism has desirable features in theory [4], such a market may not be efficient, fair and stable in practice.

In this paper, we experimentally examine a self-reporting mechanism similar to the feedback used by eBay. Effective experimental study of reputation mechanisms requires experiments long enough for behavior to stabilize. We found laboratory experiments taking place within a few hours can provide enough transaction history to distinguish "good" from "bad" behaviors and allow identifying aggregate effects on the market [5]. Our experiments include noise, which models, for example, the situation in a single transaction where the intention to pay on time cannot be distinguished, if delayed by the mail, from the intention to pay late.

We found that the self-reporting mechanism successfully maintained high fulfillment rates and we provide an analysis of individual behavior in the use of the mechanism. In particular, the reports show a high level of accuracy (95%). Furthermore, most of the inaccurate reports were instances where a player gave a bad rating for someone who *successfully* completed the transaction. This points to more subtle underlying behavior that warrants additional scrutiny. In particular, we speculate that a small percentage of the reports (5%) were used strategically to punish another player or to deliberately lower their reputation.

After describing prior related experimental studies of reputation, Sec. 3 presents our experimental setup. We then discuss the experimental results, both in terms of the choices individuals made on whether to fulfill their contracts and their effect on overall market efficiency.

2 Reputation Mechanism Experiments

A number of experimental studies have addressed the performance of various reputation mechanisms. In one approach [6], participants face an abstracted version of the transaction, namely the "trust game" where one player can choose to send money to a second, this amount is then substantially increased and the second player can choose to share some of that gain with the first player. By removing many of the complexities involved in market transactions, this game provides a simple context to study the effect of different information policies about revealing past behaviors. Addressing market efficiency requires more complex experimental scenarios [7, 3].

In contrast to this work, our experiments provide a broader set of endogenous choices for the players. First, the players can explicitly decide who they wish to do business. Although not studied in this paper, this feature allows examining whether people choose to use reputation information to ostracize those with low reputations or give them poor prices based on their higher perceived risk. Second, both buyers and sellers make delivery choices and so face a moral hazard for which reputations are relevant. In the context of a reputation mechanism based on self-reported information, this setup for reputation on both sides of a market allows players to misreport their experience as possible punishment for a poor report on their own reputation. More generally, our setup allows for the formation of clusters of mutually high-reputation trading arrangements. Third, our experiments include a full market so prices and trading volumes are determined endogenously, providing a broader view of the macroeconomic consequences of different information policies than is possible in more restricted scenarios.

3 Experimental Design

Reputation mechanisms could have complicated effects on markets. Our experiments were designed to evaluate both aggregate and individual behaviors. In particular, the experiments provide information on how people use self-reporting mechanisms and respond to unfulfilled contracts. Our experiments had two essential components: an exchange economy and an information policy for revealing past behaviors to participants. In the remainder of this section, we describe each of these in turn. All subjects received web-based instructions[1]. Each participant had to qualify by successfully passing a web-based quiz before participating in the experiment.

3.1 Exchange Economy

The first component of our experiment was an exchange economy of a single homogenous good. We used standard experimental techniques to create the market [8]. Supply and demand were generated by methods of induced value and induced cost. That is, each unit of good a buyer purchased was redeemed for a pre-determined amount, specified in an experimental currency with an announced rate at which it would be exchanged into dollars at the end of the experiment. Similarly, each unit of good a seller sold cost a pre-determined amount. Thus a buyer could profit by purchasing a unit below its redemption value, and a seller could profit from a sale above the unit's cost.

An experiment consisted of a number of periods. In each period, buyers and sellers received tables listing their redemption values and costs, respectively. The aggregate supply and demand was kept constant across periods. This fact was

[1] available at http://www.hpl.hp.com/econexperiment/marketinfo-base/instructions.htm

publicly announced at the beginning of each experiment. However, each redemption value on the demand curve and each cost on the supply curve was assigned to a random individual in each period. Thus, although the aggregate supply and demand did not change, an individual's supply and demand did change. The primary reason for this design feature is to prevent subjects learning each other's supply and demand and using this information to augment reputation information. For example, if I know that seller A always has only 3 units to sell at a cost below the specified price, I can deduce his intention to not fulfill if he offers 4 units for sale. We would like the subjects to make that determination solely based on the information provided by a controlled reputation mechanism.

We used a discrete form of double auction as the market institution as opposed to the more common continuous time version which allows a subject to submit an offer or accept an offer at any time as long as the market is open [8]. Each period consisted of a fixed number of rounds. Buyers and sellers took turns making offers (setting a price and a quantity) and accepting offers made by others. We allowed players to have only one offer at a time, although they could offer to buy or sell multiple units. There are two reasons for this form of market. First, a discrete time, round-based, design gives subjects more time between decisions to study and use information relevant to reputation compared to a continuous time version in which they may only have seconds to make a decision. Second, subjects needed to be able to choose who they would do business with. This choice was more natural in a double auction setting than a call market or other institution with a central clearing mechanism. To this end, we allowed the subjects to add a filter to their offer limiting who was permitted to accept it. Each participant could accept as many offers as were available to them. Subjects were able to see all offers, including those they were not permitted to accept. We provided this information to speed up the price formation process. When an offer was accepted, it became a contract – an agreement for the seller to produce and send the goods and for the buyer to send payment.

The key feature of our experiment was that contracts were not binding. After the last round of exchanging offers in a period, both buyers and sellers were given a list of the contracts they had signed for that period. They then decided whether or not to fulfill each contract. That is, buyers chose whether or not to send payment, and sellers chose whether or not to send the goods promised. This created an environment similar to online transactions between anonymous parties when there was no contract enforcement mechanism. Participants who chose not to fulfill their contracts avoided the associated cost of fulfillment (i.e., the payment in the case of a buyer, and the production cost in the case of a seller).

The experiment included noise: a fixed probability that either payment or goods would be lost "in transit". This probability was announced to the participants in advance. When this probabilistic loss occurred, the sending party was notified that their part of the exchange was not delivered to the recipient. However, the recipient received no such notification. Thus, for instance, a seller not receiving the contracted payment from a buyer would not know whether the buyer chose not to pay or whether the payment was lost.

3.2 Information Policies

The second component of our experiment was the information policy. This controlled the information available to subjects when they made trading and contract decisions. The focus of this series of experiment was the effect of past transaction information.

The past transaction information available to subjects varied with the experiment. Five treatments were conducted with different combinations of information policies and noise, as listed in Table 1. In each treatment, all information was displayed by period, with one row on a spreadsheet representing a period. Totals were given on a separate row. Market price (the average price of accepted contracts, weighted by volume), market volume, and personal payoffs were given in all treatments. The information policies were:

– Low information: Agents were given historical information about only their own transactions. Buyers were given the total value (the sum of price times quantity) of all contracts they signed with each seller, and the value-weighted percentage of contracts that were fulfilled by the buyer and by each seller. Sellers were given analogous information. In this case, the display merely summarized information already available from that player's transactions in previous periods.
– High information: Agents were given historical information about all transactions that took place between any buyer and any seller. All agents were given the total value of contracts that each agent signed, and the value-weighted percentage that he or she fulfilled.
– Self-reported ratings: An additional stage was added after contract fulfillment in which agents rated other players for each contract signed. After receiving information about whether they received payment or goods for each contract, they were asked to give the appropriate agent a positive (+) or negative (-) rating. After all players submitted their ratings, the ratings were made public: players saw value-weighted percentages of contracts signed by a given player for which he or she received a positive rating (in addition to the total value of contracts). If all players gave positive or negative ratings if and only if their contracts were fulfilled or not, respectively, then the information available with this policy would be similar to that for the high information case.

The treatment was announced on the day of the experiment, and subjects were given complete and accurate information about the rules and nature of the game, including the probabilistic loss of payment and goods (i.e., the amount of noise).

4 Results

In this section, we describe the experiments we performed and the resulting behaviors.

4.1 Overview

Table 1. Overview of the experiments, showing for each one the number of subjects, number of periods, whether noise was added and the information policy. In the pilot experiment (number 1), the supply and demand was different from the rest.

experiment	subjects	periods	noise?	information policy
1	8	12	no	high
2	12	13	no	high
3	14	16	no	low
4	16	14	yes	low
5	16	14	yes	high
6	16	14	yes	self
7	16	16	yes	low
8	14	16	yes	self

We conducted a total of 8 experimental sessions, summarized in Table 1. The first one was a pilot experiment with 8 subjects. The rest had at least 12 subjects. The first 3 experiments had no noise. The rest of the experiments used a noise probability of 10%.

In all experiments, market prices converged reasonably well to equilibrium within 3 periods, as expected from prior studies [8]. Thus we were able to study the effect of information policy choices in the context of a rapidly equilibrating underlying market. Notice we use the term "equilibrium" loosely here.

As expected, all experiments exhibited strong end-game effects. Subjects were told when the game would end two periods ahead of time. Furthermore, they had an expectation of finishing by 5pm on the day of the experiment. Contract fulfillment decreased sharply around 4 periods before the end of an experiment. We use all of the data, including those close to the end-game, to compare information policies. We found about 10 periods in each experiment minimally affected by the end-game, providing an indication of the effects and dynamics of reputation likely to arise in the context of a long series of repeated transactions.

4.2 Fulfillment Rates

We measure aggregate contract fulfillment by the *period fulfillment rate*. To define this value, we viewed each of the contracts signed during a period as two separate transactions, the payment sent by the buyer and the goods sent by the seller, each of which could be fulfilled or not. Each contract involves a price per unit and number of units to exchange, and its value is the product of this price and number of units. The period fulfillment rate for the buyers is the ratio of the number of contracts they fulfill to the total number of contracts. We use a similar definition for the sellers. The overall fulfillment rate is the average of those for the buyers and sellers in that period. Fulfillment is not equivalent to actual delivery:

Table 2. Observed fulfillment in the experiments. Values show the average and standard deviation of the fulfillment over periods 3 to 11, inclusive. The most significant comparison is among experiments 4 through 8, since the others did not have noise or used a different supply and demand function.

experiment	information policy	average	standard deviation
1	high	57%	20%
2	high	90%	9%
3	low	85%	7%
4	low	47%	11%
5	high	83%	9%
6	self	77%	15%
7	low	64%	16%
8	self	86%	12%

as described above, even when a person decides to fulfill a contract, the payment or goods could be lost in transit due to noise and thus not delivered to the other party. Table 2 gives the observed values for our experiments. As one can see, the two self-reporting experiments resulted in fulfillment rates (77% and 86%) substantially higher the low information experiments (47% and 64%) and close to or as good as the high information experiment (86%). This is evidence that this particular reputation mechanism sustains a high fulfillment rate as well as if full transaction information were available.

4.3 Behavior of the Self-Reporting Mechanism

Table 3. Accuracy of self-reporting. For each participant in each of the two experiments, the value is the fraction of reports that match the actual behavior of the other party to the contract(s) entered into by that individual. The average accuracy in each experiment is 0.94.

expt.																
6	0.93	0.97	1.00	0.94	0.94	0.98	0.96	0.96	0.92	0.94	0.96	0.91	0.93	0.91	0.94	0.89
8	0.97	0.89	0.95	0.89	0.98	0.95	0.98	0.94	0.98	0.85	0.98	0.95	0.93	0.98		

Table 3 shows the accuracy of the reports by the individuals in our two experiments involving self-reporting. We see the accuracy is quite high, and consistently for all participants. As a further detail on the use of the reports, Table 4 show how the ratings compare with the actual behavior. In most of the cases, the rating (good or bad) correctly corresponds to actual experience with the contract (the good or payment received or not, respectively). In the remaining cases, we see people are much more likely to give a bad report in spite

Table 4. Comparison of ratings and actual behaviors for all periods and participants in the experiments using self-reporting.

	experiment 6	
rating	received	not received
good	0.544	0.029
bad	0.083	0.344
	experiment 8	
rating	received	not received
good	0.612	0.003
bad	0.123	0.262

of receiving the contracted good or payment, than they are to give a good rating in spite of not receiving it. This bias toward bad reporting could reflect a desire to punish an individual for poor past behavior (including negative ratings) or a desire to lower the reputation of potential competition. In either case, more scrutiny to the data may shed some light on how the self-reporting mechanism was used strategically.

Furthermore, due to the noise introduced in the experiments, not receiving the contracted value does not mean the person intentionally did not fulfill: the item could have been lost due to noise.

4.4 Individual Behavior

We are interested in how individuals respond to fulfillment. Do they screen out those perceived as unlikely to fulfill contracts? Or do they enter into contracts with such people but at less favorable prices? Do they use a tit-for-tat strategy to encourage fulfillment?

We found subjects tend not to fulfill contracts with the people who didn't fulfill prior contracts. There is strong evidence that people are engaging in tit-for-tat strategies. Specifically, we analyze each individual in the following way. For each individual, we compute the total value of of his or her contracts that were not fulfilled by the other party. This serves as a naive value of how much that player distrusts a potential trader. We regress this value on the percentage of contracts that this person chooses to fulfill to the same trader. In all 78 subjects (of experiment 4 through 8), this regression yields a negative coefficient. That is, a subject tends not to fulfill his contract with another trader if this trader has not fulfilled a prior contract with him or her before. Furthermore, in 64 out of the 78 subjects (82%), this coefficient is significant

5 Conclusion and Future Work

We described a series of experiments studying the effect of reputation mechanisms. Three different mechanisms were tested: low information policy when subjects observed records of their own transactions, high information when aggregate statistics of all transactions were common knowledge and a self-reporting mechanism in which subjects scored their partners after each transaction. Results show that an ebay-style self-reporting mechanism was largely effective in preventing ”cheating” and the reports were highly accurate (95%). However, we also observe a heavy bias in the inaccurate reports towards giving ”bad” ratings to players who successfully fulfilled their contracts. It is likely that these inaccurate reports are intentional and were used strategically to either punish past transgression or to lower the reputation of the competition. This will be one interesting issue to examine in the future.

This experimental design can be easily expanded to examine a wide range of issues. For instance, with respect to self-reporting we could also compare centralized (e.g., eBay) to distributed (e.g., word of mouth recommendations) reputation mechanisms. This issue will probably depend on whether preferences are homogenous or not. For example, every seller prefers the buyer to pay promptly. Thus reputation based on paying behavior can have a common measurement. However, reputation about recommending movies will vary with the preferences of the person.

There is a direct equivalence between fulfillment as used in our experiments and that on eBay: namely not sending goods or payment. However, “fulfillment” can be interpreted to have continuous values on eBay as opposed to the way it was set up in the experiment. For example, a seller can send sub-standard goods to the buyer while advertising for perfection, or buyers could delay, but still ultimately deliver, payment. Our experimental design could be extended to treat this question by allowing sellers to choose the quality of good to produce (with different costs and benefits to the buyer) and thereby have the option of sending a lower quality good than specified in the contract with the buyer. Identification policy is another interesting issue our experimental setup could study. This would broaden the experiment to examine issues of anonymity, the ability to change identity at will and after markets of buying, selling and/or leasing identity in similar set ups. This last option, involving markets for reputation, has been studied theoretically [9] and could be readily added to our experiment design by allowing players to trade identities, and their associated transaction histories.

Privacy is another issue closely linked with reputation mechanism design. Disclosure of personal information may facilitate the establishment of one’s reputation. For example, eBay requires an email address. Obviously if an address is required and there are ways to track down a trading partner, it is easier to establish trust. However, people may prefer to keep this information private, even if they incur some cost due to less trust on the part of others. Such concerns could be addressed without need for trusted third parties through the use of cryptographic protocols [10].

By combining markets with various information policies, we were able to study the effect of reputation mechanisms under controlled laboratory conditions. Even within the limited time available for such experiments, our design allowed us to observe differences in behavior due to the amount of past transaction information revealed to participants. Moreover, our experiment design readily extends to address a variety of interesting questions beyond those described here, such as changing or trading identities. These experiments complement larger, but less controlled, field studies of reputation in practice, such as used by eBay, and theoretical studies relying on simplifying assumptions of rational behavior or limited to deal with analytically tractable games.

References

1. Klein, D.B., ed.: Reputation: Studies in the Voluntary Elicitation of Good Conduct. Univ. of Michigan Press, Ann Arbor (1997)
2. Axelrod, R., Hamilton, W.: The evolution of cooperation. Science **211** (1981) 1390–1396
3. Dejong, D.V., Forsythe, R., Lundholm, R.J.: Ripoffs, lemons and reputation formation in agency relationships: A laboratory market study. J. of Finance **XL** (1985) 809–820
4. Dellarocas, C.: Analyzing the economic efficiency of eBay-like online reputation reporting mechanisms. In: Proc. of the 3rd ACM Conf. on Electronic Commerce (EC01). (2001) 171–179
5. Chen, K.Y., Hogg, T., Wozny, N.: Experimental study of market reputation mechanisms. In: Proc. of the 5th ACM Conference on Electronic Commerce (EC'04), ACM Press (2004) 234–235
6. Keser, C.: Experimental games for the design of reputation management systems. IBM Systems Journal **42** (2003) 498–506
7. Bolton, G.E., Katok, E., Ockenfels, A.: How effective are online reputation mechanisms? Technical report (2002)
8. Smith, V.L.: Bargaining and Market Behavior: Essays in Experimental Economics. Cambridge Univ. Press (2000)
9. Tadelis, S.: The market for reputations as an incentive mechanism. SIEPR Policy Paper 01-001, Stanford (2001)
10. Huberman, B.A., Franklin, M., Hogg, T.: Enhancing privacy and trust in electronic communities. In: Proc. of the ACM Conference on Electronic Commerce (EC99), NY, ACM Press (1999) 78–86