Experimental Study of Reputation Mechanisms in an Exchange Economy

Kay-Yut Chen, Tad Hogg
Hewlett-Packard Laboratories

Nathan Wozny
California Institute of Technology

May 28, 2004

Abstract

We experimentally evaluate reputation mechanisms in an exchange market in which participants have the option of not fulfilling their contracts. These mechanisms vary in the information they provide on past behavior. Participants can choose who they trade with, allowing endogenous response to low reputation through ostracizing or price discriminating behaviors. The participants responded strategically to the provided information. In particular, mechanisms revealing more information had a statistically significant increase in fulfillment rates. Our experimental design allowed identifying the effect of reputation mechanism on endogenous market behavior. We found that decreasing fulfillment led to lower efficiency and increased market volume but did not affect the prices.

1 Introduction

Reputation has been an important aspect of commerce since the emergence of exchange economies [7]. 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.

More generally, reputation can help address two issues: moral hazard and adverse selection. In this paper, we focus on moral hazard: whether a party to a contract will choose to fulfill their part of the contract (sending payment or delivering the good, for a buyer or seller, respectively). Adverse selection involves people with hidden heterogenous exogenous characteristics such as degrees of competence at providing a good or service. Since these characteristics are hidden to potential customers beforehand, they may question whether the advertised terms of a contract accurately reflect the quality they will receive. As a result, a proper reputation mechanism could also help address adverse selection, i.e., distinguishing among sellers with different characteristics. Although this is also an interesting issue, the experiments described in this paper focus on moral hazard. That is, we focus on situations in which people face the question of whether a proposed transaction will proceed as agreed. This is a significant issue in the context of the internet because although a given person may engage in many transactions, only a few may be with a particular person, thereby limiting the ability to learn from direct experience.

Establishing trust through repeated interactions has been studied in several contexts, particularly with the iterated Prisoner's Dilemma [1]. 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.

Reputation mechanisms in the context of the Internet should consider the following factors:

- There are many small players with no pre-established reputation.
- External enforcement mechanisms are not available (e.g., due to the difficulty of connecting an on-line identity with a physical person) or expensive to use compared to the value of the transaction (e.g., due to jurisdictional ambiguity).
- Identity is anonymous. Thus one can "reset" to a zero reputation at

anytime.

- Identities and their associated reputations may themselves be traded, and it may be difficult for others to notice such a change in ownership, at least for a while.

To focus our experiments, we examine markets without contract enforcement, incorporating the first two of these factors. Nevertheless, the experimental design reported in this paper readily extends to allow changing identities of the participants, either as individual choices to restart with a new identity or through trades of existing identities and their associated reputation histories. Thus our design allows separate examination various combinations of these factors relevant to reputation for small transactions via the Internet.

In this paper, we experimentally examine the behavior of mechanisms revealing differing amounts of information on the transaction history of the players. These mechanisms offer participants strategic choices and incentives that are difficult to model precisely. In particular, game theory does not accurately predict actual behavior and, in many cases, the mechanisms are too complicated to allow analytic evaluation of the game theory predictions.

Effective experimental study of reputation mechanisms requires experiments long enough for behavior to stabilize. In particular, participants need to build up enough transaction history to distinguish "good" from "bad" behaviors. Furthermore, the value of reputation declines toward the end of an experiment since there are fewer future opportunities for trades. Even using a probabilistic ending time would not remove this problem because subjects know they will be leaving by a certain time of the day. On the other hand, observations close to the end of experiments show how markets behave when reputation mechanisms break down. Thus an important empirical question is the extent to which laboratory experiments are feasible when they include both a reputation mechanism and the time required for an underlying exchange market. Including both features is desirable in allowing participants to endogenously make choices in the market and hence revealing the effect of reputation on aggregate market behavior. Such observations are not possible in experiments that focus solely on simple cooperate/defect choices (e.g., as in a study of the iterated Prisoner's Dilemma).

Beyond these questions of the feasibility of the experiments, our main research focus in this paper is on aggregate behaviors affected by the mechanisms, e.g., how market efficiency, equilibrium price and traded volume vary with the choice of mechanism. We also examine noise tolerance in the mechanisms. This noise 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.

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 Experiments on Reputation Mechanisms

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.

Questions involving reputation's effect on market efficiency require more complex experimental scenarios. One example is a fixed-price market in which sellers have the option of not fulfilling their contracts [2]. Players are paired by the experimenter, and randomly assigned roles of buyer or seller. The buyer chooses whether or send the purchase price to the seller and, if so, the seller then decides whether to deliver the good. Revealing the seller's history of fulfillment then provides reputation information for the buyer.

A similar one-sided market, but allowing variable prices, arises in an experimental study of a principal/agent market [3]. In this case the agents ("sellers") choose the level of service to provide after principals ("buyers") have committed to their contracts. In this experiment, the agents face a market choice, maximizing the difference between their fee and expense of providing their service, but the payoff for the principal is probabilistic rather than directly determined by the choice made by the agent.

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 with rather than being paired with a single other player by the experimenter. 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 answer three questions about this interaction.

First, how does reputation affect the level of fulfillment in a market, where participants endogenously set prices and select their business partners? For example, experience with eBay suggests a self-reporting mechanism encourages cooperation. Since we cannot rerun the eBay auctions *without* its feedback mechanism, this example does not provided direct evidence of the difference it has made. Of course, intuition suggests eBay would have collapsed without some kind of reputation mechanism. Our experiments allowed us to measure the effect of such a reputation mechanism in a market.

Second, does the option to not fulfill contracts significantly change the macroeconomic behavior of the market? In particular, how does it affect the equilibrium price and volume, and the resulting efficiency. Addressing this question within the time available for a laboratory experiment requires an underlying market mechanism that rapidly converges to equilibrium even with relatively few participants. For example, many experimental studies [9] have shown that double-auction markets, with simple induced supply and demand curves, converge to the competitive equilibrium within a few periods even with only a small group of subjects (e.g., about 10). With these observations in mind, we use double-auction markets as the basis of our experiments.

Third, how do individuals behave in response to unfulfilled contracts? One possibility is ostracism: individuals refusing to do future business with those who did not fulfill in the past. Another possible reaction is price discrimination: offering better terms to those with a history of fulfillment. Our experimental design allowed participants, on both sides of the market, to select potential business partners, thus allowing them to endogenously select among these possibilities.

In this paper, we focus on the first two questions, the aggregate effects of different reputation mechanisms. 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 [9]. 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

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

sellers received tables listing their redemption values and costs, respectively. The aggregate supply and demand was kept constant across periods. This fact was 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 [9]. 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. It was also possible to vary how subjects' identities were managed but we did not do so in these experiments. Instead, we used one identification policy: each subject received a single unique id for the entire experiment.

The past transaction information available to subjects varied with the experiment. There were two treatment variables: information policy and noise. 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.

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

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.

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

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. Rapid convergence is expected for the market portion of the experiment from prior studies [9]. 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. We did observe quick convergence of the price to the price where the supply and demand curve crossed. However, since we lack a complete theory of this scenario, we do not know whether that is a competitive market equilibrium with the addition of nonfulfillment possibilities and reputation mechanisms.

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

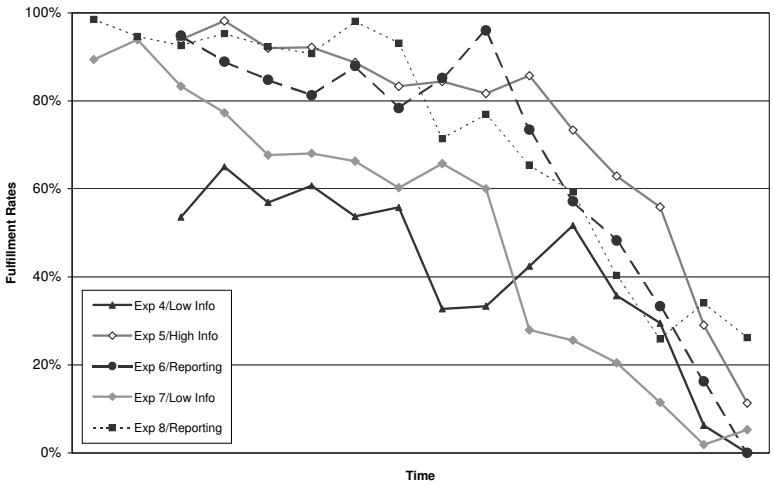

Figure 1: Fulfillment rates vs. time for the experiments with noise (numbers 4 to 8). All curves are aligned at the last period of the corresponding experiments.

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 Information Policies and Fulfillment

We used the *period fulfillment rate* as the key measurement of the aggregate level of contract fulfillment in the market. 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: 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.

The fulfillment rate varies both within an experiment (e.g., due to the end-game effect) and between experiments (e.g., due to different information policies). Thus a direct comparison of differences in aggregate fulfillment rates (in Table 2) is not statistically significant due to the large variation within each experiment. On the other hand, determining whether mean fulfillment rates differ by comparing fulfillment on a period-by-period basis with the two-population

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%

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.

	2	3	4	5	6	7	8
2	–	0.92	1	0.92	0.83	1	0.77
3	0.08	–	0.92	0.29	0.46	0.81	0.38
4	0	0.08	–	0	0	0.38	0.08
5	0.08	0.71	1	–	0.77	1	0.64
6	0.17	0.54	1	0.23	–	1	0.46
7	0	0.19	0.62	0	0	–	0
8	0.23	0.62	0.92	0.36	0.54	1	–

Table 3: For each pair of experiments, the fraction of compared periods in which the experiment listed in a row had higher fulfillment than the experiment listed in a column. For example, the value in the first row and second column is the fraction of periods with higher fulfillment in experiment 2 than in experiment 3.

	2	3	4	5	6	7	8
2	–	$\mathbf{2 \times 10^{-3}}$	$\mathbf{2 \times 10^{-4}}$	$\mathbf{2 \times 10^{-3}}$	**0.02**	$\mathbf{10^{-4}}$	**0.05**
3	1	–	$\mathbf{2 \times 10^{-3}}$	0.97	0.71	**0.01**	0.89
4	1	1	–	1	1	0.87	1
5	1	0.09	$\mathbf{10^{-4}}$	–	**0.05**	$\mathbf{6 \times 10^{-5}}$	0.21
6	1	0.5	$\mathbf{10^{-4}}$	0.99	–	$\mathbf{10^{-4}}$	0.71
7	1	1	0.29	1	1	–	1
8	0.99	0.23	$\mathbf{2 \times 10^{-3}}$	0.91	0.5	$\mathbf{2 \times 10^{-5}}$	–

Table 4: For each pair of experiments, the probability, under the null hypothesis, of obtaining at least as many periods with higher fulfillment in the first (row) than in the second (column) as observed experimentally. Entries in bold indicate cases where the experiment listed in the row gave statistically significant (at the 0.05 level) higher fulfillment than the experiment listed in the column.

t-test is inappropriate because periods in a single experiment are not independent.

Instead, as a simple way to remove the variation within an experiment, we compare periods in two experiments starting from the last period of each. Specifically, consider two experiments A and B with T_A and T_B periods, respectively. For $i = 0$ up to $\min(T_A, T_B) - 1$, we compare the fulfillment rates in periods $T_A - i$ and $T_B - i$ of experiments A and B, respectively, and count the fraction of times the fulfillment rate in A is larger than that in B. Matching periods in two experiments by the number of periods remaining until the end seems reasonable because the end game effect in all experiments appeared to be about the same, independent of the total number of periods in the experiment, and was the dominant variation in behavior during an experiment. The result of this pairwise comparison is shown in Table 3.

This comparison ignores the *amount* by which fulfillment differs in the two experiments, but allows a simple statistical comparison with the null hypothesis that the two experiments have the same behavior with respect to fulfillment without requiring assumptions about how the end game effect changes fulfillment during an experiment. Specifically, under the null hypothesis the probability experiment A had higher fulfillment rate in each comparison is 50%. Thus the number of times experiment A had higher fulfillment rates than B is distributed according to the binomial distribution with probability 1/2, i.e., $B(n, k) \equiv \binom{n}{k} 2^{-n}$ is the probability to have exactly k out of n comparisons give higher values for experiment A. When we experimentally observe K periods for which experiment A has higher fulfillment, the null hypothesis gives at least this many times with probability $\sum_{k \geq K} B(n, k)$. This probability is the p-value reported in Table 4.

For example, in comparing experiments 5 (high information) and 8 (self-reporting), Table 3 shows experiment 5 had higher fulfillment than experiment 8 in 9 of the last 14 periods of the experiments, corresponding to the value of $9/14 = 64\%$ in the table. Under the null hypothesis that these two experiments had the same behavior with respect to fulfillment, we would expect to see 9 or more periods with higher fulfillment in experiment 5 than in 8 with probability $\sum_{k \geq 9} B(14, k) = 0.21$, as shown in Table 4.

4.2.1 Less information reduces fulfillment

Our first observation from Table 2 is the lower fulfillment rates in experiments 4 and 7 than in the others. Specifically, when comparing experiment 5,6 and 8 (high and self-reporting information policies) to experiment 4 and 7 (low information), the null hypothesis was rejected in favor of the alternative that experiment 5,6 and 8 had higher fulfillment rate with very high significance. In fact, as seen in Table 3, experiments 5,6 and 8 had higher fulfillment than 4 and 7 in all periods except one. The p-value of a case when the fulfillment rates of one experiment dominated the other one completely at 100% is $(1/2)^n$ where n is the number of observations. In the case of comparing experiment 4 and 5 (with 14 observations), the p-value is 6.1×10^{-5}.

The same test comparing the two low-information experiments (4 and 7) cannot reject the hypothesis that fulfillment rate in either experiment is equally likely to be higher than the other.

This comparison is consistent with intuition as well as a theory that assumes the existence of some irrational players. Under this class of models, rational players have the incentive to mimic the irrational players in response to the information policy. In any case, it is reasonable to expect people are more likely to fulfill contracts when they know more information about their actions is available to other participants. In the low information policy treatment, subjects can safely reason that only people whom they have failed to fulfill know of their "bad" record. Thus, fulfillment is likely to be less important for future business for low information than with full information. Our observations are also strong evidence that people react to strategic implications of how information can and will be used by others.

Interestingly, in a small market (with around 16 subjects) even information limited to an individual's own transactions (the low information policy) supports a significant amount of cooperation.

4.2.2 Self-reporting is as good as perfect information

A second observation from Table 3 is the fulfillment rates in experiment 6 and 8, which use self-reporting, are indistinguishable from experiments with a high information policy. Specifically, Table 4 indicates experiment 5 (high information) had higher fulfillment rates than experiment 6 (self-reporting) at 5% significance. However, experiment 5 did not have higher fulfillment rates when compared to experiment 8. The test performed to compare experiment 6 and 8 cannot reject the hypothesis that fulfillment rate in either experiment is equally likely to be higher than the other. It would seem that self-reporting was almost as good as high information but not quite.

This finding is consistent with observations on Internet auction sites such as eBay. On eBay, the amount of negative feedback is extremely small, on the order of 1% of the total amount of feedback [8]. This low value of negative feedback is not direct evidence that fulfillment is high since feedback does not have to be 100% accurate. However, it is fairly reasonable to deduce that lack of fulfillment is not a big problem.

Similarly, the reports in our experiments were not always accurate, as shown in Table 5. Specifically, in about 6% of the transactions the report (good or bad) did not match whether delivery occurred or not, giving the conditional probabilities

$$P(\text{good}|\text{no delivery}) = 2/(2+37) = 5\% \quad (1)$$

$$P(\text{bad}|\text{delivery}) = 4/(4+57) = 7\% \quad (2)$$

Noise may account for the good reports in spite of no delivery (quantified in Eq. (1)). Players may wish to report whether the other party fulfilled their transaction, rather than whether delivery took place. A player who receives the payment or good knows for certain that the other party chose to

	delivered?	
report	yes	no
good	57%	2%
bad	4%	37%

Table 5: Accuracy of self-reporting, from experiments 6 and 8. Values show the percentage of transactions with each type of report (good or bad) and actual outcome (whether the good or payment was delivered, to buyer and seller, respectively). Overall, 94% of the reports correspond to the actual outcome. These are from a total of 1738 transactions (869 contracts).

fulfill, but lack of delivery could either be due to noise or a choice not to fulfill. About 61% of the transactions in Table 5 were delivered, hence people could deduce about 67% of them were fulfilled since the noise probability of 10% was announced to participants. Thus among the 39% of cases not delivered, 7% came from fulfilled transactions. If people use ratings to indicate their estimates of person's intended behavior (instead of outcome), we'd then expect to see $P(\text{good}|\text{no delivery}) = 7/39 = 18\%$, quite a bit higher than the observation of only 5%. Specifically, out of the 678 transactions with no delivery, if $P(\text{good}|\text{no delivery})$ were equal to 18%, the probability we would in fact observe 5% or fewer is given from the binomial distribution as 5×10^{-24}. Hence our observations indicate participants are significantly more likely to give bad reports than one would expect if they were using reports to indicate whether the transaction was fulfilled.

We gain further insight into the appearance of bad reports in spite of delivery ("bad mouthing") from Fig. 2. There was a clear upward trend of the frequency of bad mouthing during the experiments. One possible reason is participants using reports to punish past actions. For example, suppose a seller had not fulfilled my contracts in the last several periods. Although he fulfilled this time, I could give a "bad" feedback report for the past behavior. Since the level of fulfillment generally declined through the course of the experiment, an upward trend of the "bad mouthing" phenomenon is consistent with our reasoning. By contrast, there was no clear trend in the converse behavior (good reports in spite of nondelivery).

4.2.3 Buyer and seller fulfillment differed

To compare the effect on different information policies on fulfillment, we mainly focus on the observed overall fulfillment rate, which combines behaviors of buyers and sellers. However, as seen in Fig. 3, sellers tend to have higher fulfillment rates than buyers, especially when those rates are fairly low (i.e., mainly near the end of the experiments). To quantify this observation, among the 25 periods in these experiments in which either buyer or seller fulfillment rates (or both) were at most 50%, we found 80% of these periods had higher seller fulfillment than buyer. By comparison, the null hypothesis that buyers and sellers fulfilled

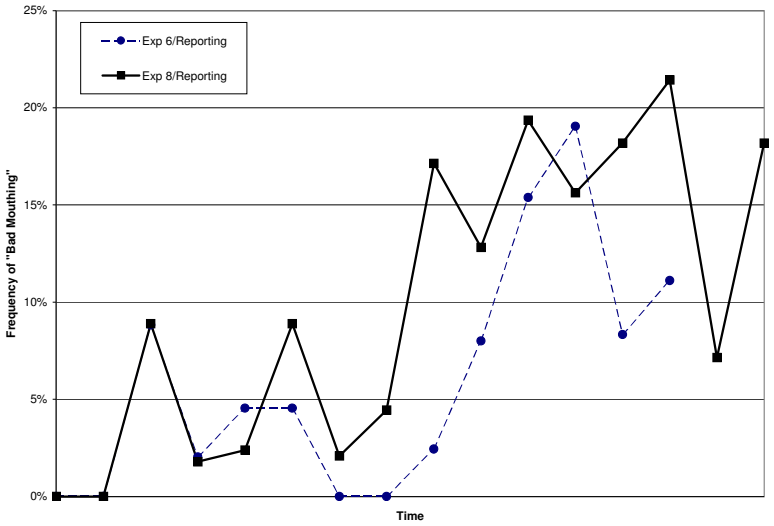

Figure 2: Frequency of "bad" reports for cases in which the good or payment was delivered, as a function of time during experiments 6 and 8.

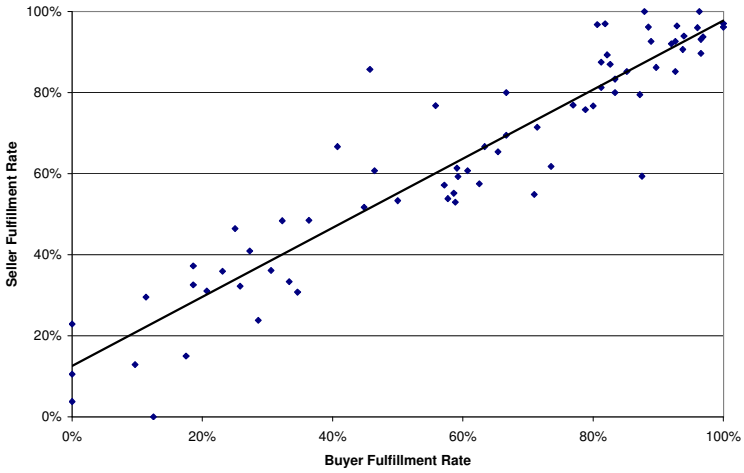

Figure 3: Fulfillment rates of buyers and sellers. Each point is the aggregate fulfillment of buyers and sellers in one period of an experiment. Points are from experiments 4 to 8.

with equal probability would give higher fulfillment by sellers at least 80% of the time would occur only with probability 2×10^{-3}. On the other hand, for periods in which both buyers and sellers had fulfillments greater than 50%, the difference between buyers and sellers was not statistically significant.

One possible explanation for this difference is that at low fulfillment rates, buyers have more to gain than sellers from not fulfilling their contracts. This is because with low fulfillment, only a few units are traded, and these are just the first few units in the aggregate supply and demand. Such units are those most valuable to the buyers and least costly for the sellers to produce. Thus, by not fulfilling a contract, a seller only saves the small production cost while a buyer saves the full payment, which is much larger since trades take place near the equilibrium price (as discussed in Sec. 4.3.2 and shown in Fig. 6). Thus given the fact that price levels remained constant, this difference in behavior of buyers and sellers was consistent with their different payoffs for not fulfilling contracts. However, this leaves open the larger question of explaining why prices did not drop in response to this difference.

4.2.4 Noise reduces fulfillment

Comparing experiment 2 with 5 (high information), and 3 with 7 (low information), we see from Table 4 that noise reduces fulfillment for both high and low information policies. A simple explanation for this observation is that with no noise, any lack of delivery immediately indicates the other party deliberately decided not to fulfill the contract. With noise, there is some doubt as to the cause. Players thus apparently adjust their behavior to take advantage of the noise.

A further comparison, low information with no noise (experiment 3) and high information with noise (experiment 5), shows no significant difference. Thus the level of noise we used caused changes in aggregate behavior comparable to the difference in information policy. The level of noise chosen was 10%, obviously higher than the actual chance of losing something in the mail. There is no particularly strong reason for this choice except to make it high enough to ensure an impact in the experiments, which have an expected length of around 15 periods, but not so high as to completely mask individual choices of whether to fulfill. Arguably 9% or 11% will not make a big difference.

4.3 Aggregate Market Behavior

We measure aggregate market behavior in three dimensions: market efficiency, price and volume. Conventional wisdom argues that the choice of whether to fulfill contracts makes the market inefficient. If an information policy (or the lack of one) reduces fulfillment, then we will expect a market with no trade or very few trades at arbitrary prices and almost zero efficiencies.

We measure market efficiency of a period by the ratio of realized surplus (i.e., the total payoff of the subjects) to the maximum surplus. In the calculation of the maximum possible surplus the participants could realize through their

choices, we modified the standard definition [9] to account for the fact that noise introduced a loss into the system beyond the control of the participants. To do this, we first assume the market clears at the original equilibrium price and volume it would have without noise, which would give the maximum possible surplus when there is no noise. Then we remove the units and payment expected to be lost due to the noise to obtain the maximum surplus adjusted for noise. Specifically, let S be the maximum surplus available from our choice of aggregate supply and demand functions, ϵ be the noise level, and P_{eq} and V_{eq} be the equilibrium price and volume (where the aggregate supply and demand curves cross). Then

$$S_{\text{avg}} = S(1-\epsilon) - \epsilon P_{\text{eq}} V_{\text{eq}} \tag{3}$$

is the expected amount of surplus if the market clears at the equilibrium price and quantity and every contract is fulfilled. The first term in this expression ($S(1-\epsilon)$) is the average remaining surplus after some produced units are lost due to noise during transit to the buyer. The second term ($\epsilon P_{\text{eq}} V_{\text{eq}}$) is the average amount of payment lost from the noise during transit to the seller. Payments or goods lost to noise are non-recoverable. Using this quantity we measure efficiency E as

$$E = \frac{P}{S_{\text{avg}}} \tag{4}$$

where P is the total payoff to all participants actually achieved in an experiment.

With this definition of efficiency, experiments with noise can realize an adjusted efficiency above 100% because S_{avg} in Eq. (3) is the *expected value* of the maximum surplus. For instance, in a lucky period when no units or payments are lost the participants could realize up to the total noise-free surplus, S, giving $E = S/S_{\text{avg}} > 1$. Thus realized surplus could be greater than the adjusted maximum surplus, which assumes an average number of units lost.

In the remainder of this section, we describe the aggregate behaviors we observed.

4.3.1 Low fulfillment reduces efficiency

Fig. 4 shows how efficiency varies with fulfillment rates in each period, pooled from all experiments. Visual inspection of the figure shows that when fulfillment is high, beyond 80%, efficiencies are close to one and insensitive to fulfillment. Furthermore, a linear regression of fulfillment rates of 80% or above on the corresponding efficiencies yields insignificant coefficient (coefficient of 0.23 with a standard error of 0.29) and a small r-squared value: $R^2 = 0.014$. This is consistent with other double auction experiments which also yielded high efficiencies [9].

By contrast, when fulfillment is low (typically near the end of each experiment) Fig. 4 shows a substantial decrease in market efficiency. Furthermore, a linear regression of fulfillment rates of 50% or less is strongly correlated with its corresponding efficiency. The slope coefficient of the linear fit is 1.16 with a standard error of 0.18, a t-value of 6.51 and $R^2 = 0.59$.

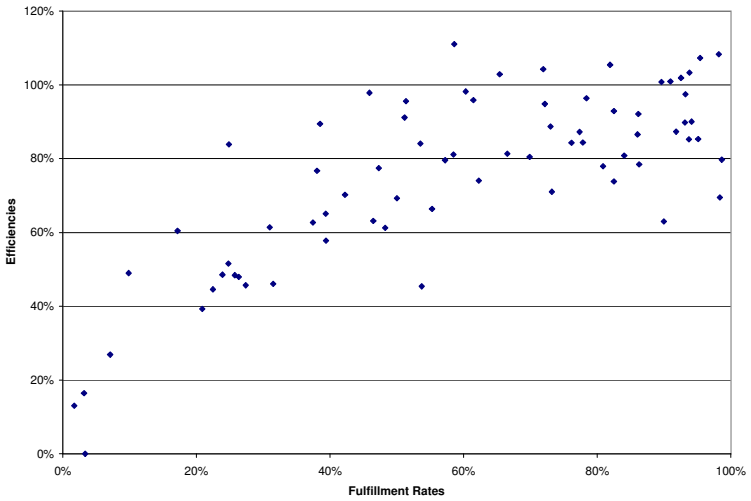

Figure 4: Market efficiency vs. fulfillment rates from all periods of the experiments with noise, i.e., numbers 4-8.

From an aggregate economic perspective, this shows that overall efficiency can tolerate a small amount of unfulfilled contracts. Thus in designing reputation mechanisms, it is sufficient if they manage to maintain high, but not necessarily perfect, fulfillment.

4.3.2 Market price and volume

Fig. 5 shows that market volume, measured as the number of units contracted, whether or not the contract is subsequently fulfilled, decreases with fulfillment. Prices, on the other hand, are fairly independent of the fulfillment rate, as seen in Fig. 6. In general, when fulfillment rates were high, the market behavior was similar to other double auction experiments in which the market converged to equilibrium prices and quantities. Explaining the behavior for low fulfillment rates is an interesting open question. In particular, in spite of the different behaviors and payoffs from non fulfilling contracts between buyers and sellers (see Sec. 4.2.3), we do not see an effect on the average price at which contracts are formed. Thus, for instance, in spite of the symmetric position of buyers and sellers in the double auction market, the prices do not adjust to bring the buyer and seller payoffs from not fulfilling closer together even though the overall supply and demand curves do not change during the experiment.

Another interesting observation from Fig. 6 is that average per-period prices are about one standard deviation below the equilibrium, even with high fulfillment.

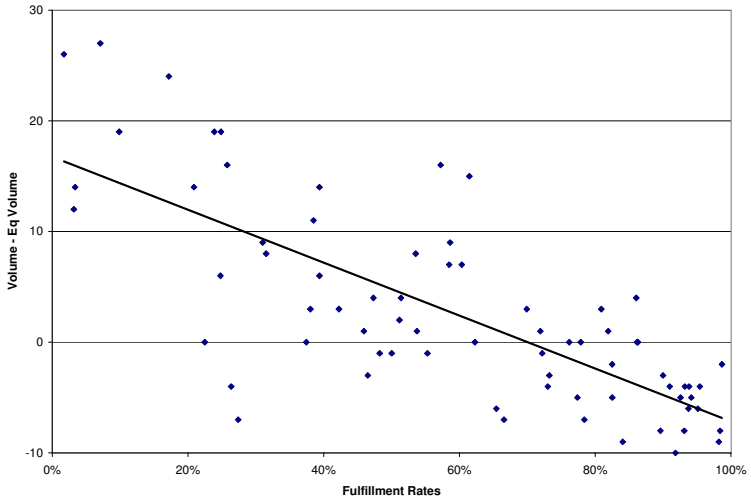

Figure 5: Market volume from all periods of the experiments with noise (numbers 4-8).

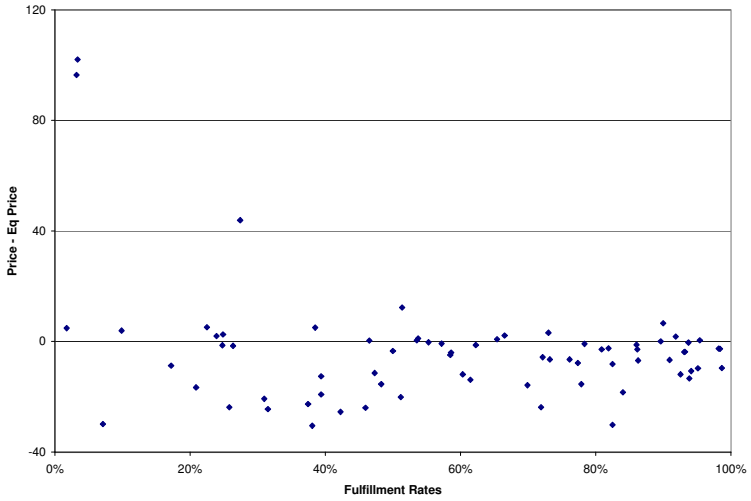

Figure 6: Market price from all periods of the experiments with noise (numbers 4-8).

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. We found fulfillment behavior responded strongly to the information policy implemented in the experiment. Thus subjects responded strategically to the effect of information about their behavior made known to the market. Furthermore, we observed some interesting market dynamics. In particular, when the amount of fulfillment decreased, the contracted volume increased but prices remained constant. Market efficiencies responded to the level of fulfillment in a highly nonlinear fashion.

The results in this paper focus on aggregate behaviors resulting from reputation mechanisms. Our experimental data also allow examining a range of questions about individual behavior, specifically how individuals respond to others with high or low reputations. In the experiment, individuals can respond to individuals with low reputation in multiple ways, e.g., avoiding future business with them, giving them less favorable pricing, being less likely to fulfill contracts with them.

This experimental design can be easily expanded to examine a wide range of issues. One direction is to study other information policies. For instance, even in the high information treatment, the fulfillment percentages are aggregates and did not allow players to distinguish whether a low fulfillment rate was with respect to all contracts or concentrated on others with especially low fulfillments, i.e., acting to punish players perceived as ignoring social expectations on behavior [11]. We can 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 rep-

utation, has been studied theoretically [10] and could be readily added to our experiment design by allowing players to trade identities, and their associated transaction histories.

One can also turn to theoretic analysis to help explain some of the results observed in the experiments. A fully rational model would conclude that people should never fulfill contracts (by backward induction) and thus the market should collapse. As one approach, we are examining a bounded-rational model of individuals with imperfect beliefs of the future value of their reputation.

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 [5].

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] R. Axelrod and W. Hamilton. The evolution of cooperation. *Science*, 211:1390–1396, March 1981.

[2] Gary E. Bolton, Elena Katok, and Axel Ockenfels. How effective are online reputation mechanisms? Technical report, 2002.

[3] Douglas V. Dejong, Robert Forsythe, and Russell J. Lundholm. Ripoffs, lemons and reputation formation in agency relationships: A laboratory market study. *J. of Finance*, XL:809–820, 1985.

[4] C. Dellarocas. Analyzing the economic efficiency of eBay-like online reputation reporting mechanisms. In *Proc. of the 3rd ACM Conf. on Electronic Commerce (EC01)*, pages 171–179, 2001.

[5] Bernardo A. Huberman, Matt Franklin, and Tad Hogg. Enhancing privacy and trust in electronic communities. In *Proc. of the ACM Conference on Electronic Commerce (EC99)*, pages 78–86, NY, 1999. ACM Press.

[6] Claudia Keser. Experimental games for the design of reputation management systems. *IBM Systems Journal*, 42(3):498–506, 2003.

[7] Daniel B. Klein, editor. *Reputation: Studies in the Voluntary Elicitation of Good Conduct*. Univ. of Michigan Press, Ann Arbor, 1997.

[8] Paul Resnick and Richard Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system. In Michael R. Baye, editor, *The Economics of the Internet and E-Commerce*, volume 11 of *Advances in Applied Microeconomics*. Elsevier, Amsterdam, 2002.

[9] Vernon L. Smith. *Bargaining and Market Behavior: Essays in Experimental Economics*. Cambridge Univ. Press, 2000.

[10] Steven Tadelis. The market for reputations as an incentive mechanism. SIEPR Policy Paper 01-001, Stanford, 2001.

[11] Masaru Tomita and Takashi Kido. Sacrificial acts in single round prisoner's dilemma. *Informatica*, 18:411–416, 1994.