Chapter 1

The Limits of Trust in Economic Transactions: Investigations of Perfect Reputation Systems

Gary E. Bolton and Axel Ockenfels

As the Internet economy has grown, so too has the need for trust. A degree of trust is critical in virtually all economic relationships, Internet or otherwise. Every day we choose to trust plumbers, doctors, employers, employees, teachers, airlines, and others. The need for trust arises from the fact that we cannot contract on every move others make. And what we can contract on is often prohibitively costly to enforce. The anonymity of geographically dispersed Internet traders increases contracting difficulties: you may not be able to identify your eBay seller or verify the quality of the object being sold, let alone get your money back.[1]

The economic foundation of trust relationships is the reciprocity principle of tit-for-tat combined with reputation systems that store information on past performance (Greif 1993). Broadly speaking, there are two forms. *Direct* reciprocity applies to repeated relationships: 'I will trust you tomorrow if you are trustworthy with me today,' and is associated with bilateral reputation systems. *Indirect* reciprocal systems enforce trust when the relationship is one-shot by a more circuitous tit-for-tat: 'I will trust you tomorrow if you are trustworthy with a third party today,' and is associated with multilateral reputation systems. Internet markets tend to be anonymous places and feature a lot of one-time transactions. A study by Paul Resnick and Richard Zeckhauser, for example, found

that a large majority of eBay trading encounters are one-shot (2002). As a result, Internet markets tend to lean heavily on multilateral systems to enforce trustworthiness (for a discussion of how community relationships influence the control mechanisms the community will accept, see chapter 9, this volume).

Take eBay's famous feedback forum, a kind of institutionalized gossip.[2] On eBay, after each encounter, buyers and sellers can evaluate each other by giving one another either a positive (+1), neutral (0) or negative (–1) feedback score, and maybe additional commentary. This feedback is publicly available and easy to access, so that each buyer can look at a seller's feedback history before he engages in bidding. The incentives for moral hazard are thus weakened by the feedback system: if traders punish sellers with negative feedback by refusing to buy from them or reducing the price they are willing to pay, then the threat of leaving negative feedback should discipline the seller.

In this chapter, we discuss our investigations of perfect reputation systems for indirect reciprocity. By *perfect*, we mean that the information about traders' past behavior circulating through the market is comprehensive and reliable. Of course, real world reputation systems are imperfect. By studying perfect reputation systems, however, we identify the maximal achievable benefit by market design improvements—absent from all kinds of institutional noise and incentive problems inherent to real world reputation systems. By the same token, we reveal how behavioral aspects may limit or assist the reputation system performance, and we get a clearer measure of the interplay of institutional and behavioral aspects on the effectiveness of reputation systems.

Studying perfect reputation systems is an important complement to field studies of feedback systems such as eBay's. One reason is that though most of the empirical literature observes that traders respond to reputation information, this observation in itself does not measure the virtues of reputation information. Evidence from eBay, for instance, shows that a seller's feedback profile may affect prices and the probability of sale (see Dellarocas et al. 2004; Dellarocas 2006; Resnick et al. 2006). The empirical results are mostly consistent with the theoretical expectation of buyers paying more to sellers with better reputations. It has also been observed that the impact of reputation ratings on buyer behavior tends to be stronger for riskier transactions and more expensive objects. This would seem to indicate that the reputation systems have at least some merit. But precisely how much is gained from these systems in terms of overall cooperation levels and efficiency gains remains unclear. We can get a sense of this measure by studying perfect systems under laboratory conditions. A second reason is that field studies have difficulties separating imperfect institutions and boundedly rational behavior. It may be that flawed systems work well because real-world

traders cannot exploit the flaws as fully as theories assuming full rationality would suggest they do. Current studies are discovering changes in rules, procedures, and information aggregation that may well help generate more reliable information. Retaliatory feedback might be eliminated by not letting sellers evaluate buyers, as suggested by Werner Güth, Friederike Mengel, and Axel Ockenfels (2006), or by having a blind period in which trading partners can simultaneously leave feedback on each other, as suggested by Tobias Klein and colleagues (2007). Clever incentive schemes, based on economics (Miller, Resnick, and Zeckhauser 2005) or social psychology (Rashid et al. 2006), may overcome the public goods problem and promote full provision of all relevant feedback information. Modern authentication technologies or entry fees may also eliminate manipulative changes of online identities (see Friedman and Resnick 2001; Ockenfels 2003). But maybe the binding limitation for the effectiveness of reputation systems is not so much the institutional issues but rather the behavioral limitations. Studying perfect systems can cleanly expose these kinds of limitations.

It is also important to recognize that some of the pressing challenges that the imperfection of information in real-world systems creates have to do with strategic behavior (see also Bolton, Greiner, and Ockenfels 2008). One challenge for feedback systems such as eBay's is that feedback information must come from voluntary self-reporting of one's experiences with trading partners. Feedback is a public good, however; the costs of providing feedback are paid by the provider but the benefit goes only to other traders.[3] Furthermore, no trader can be excluded from using the information. As a result, economic theory suggests that feedback information will be underprovided. In fact, only about 50 to 70 percent of the transactions on eBay receive feedback (Resnick and Zeckhauser 2002; Bolton, Greiner, and Ockenfels 2008).[4] A second challenge is that feedback needs to be reliable to effectively deter fraudulent behavior. There are a variety of incentives to manipulate feedback, for example, to give good feedback to friends and bad feedback to competitors. A third major challenge is that negative feedback is often retaliated by additional negative feedback, creating incentives to not give negative feedback. It appears suspicious that less than half percent of the eBay feedback is negative (as observed by, among others, Resnick and Zeckhauser 2002). Further evidence for the limited reliability of eBay's feedback information comes from the observation that negative feedbacks are given late, in the last minute. On the other hand, positive feedback tends to be given earlier, to trigger a reciprocal response (for example, Klein et al. 2007; Bolton, Greiner, and Ockenfels 2008). As a consequence, the information value of feedback, if given at all, is likely to be something less than perfect.[5] A better understanding of the strategies that people pursue in a perfect, idealized system can help us identify

and understand the strategies they pursue in more complex environments.

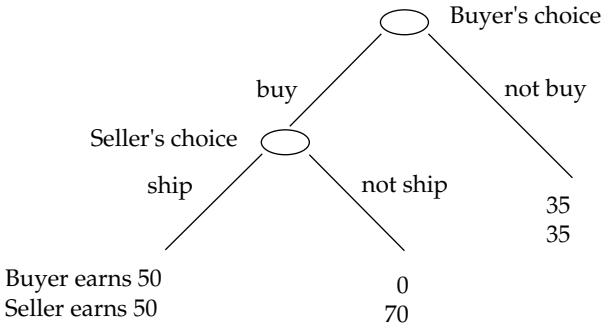
We study the scope and limitations of perfect reputation systems in thought experiments, using economic theory, and in laboratory experiments, exposing people to perfect systems. What we find, as we illustrate, is that economic theory tends to underestimate traders' intrinsic willingness to behave reciprocally, but at the same time to overestimate the effectiveness of extrinsic motivations induced by reputation institutions. One implication of our work is that understanding how social behavior can be sustained with the help of reputation mechanisms will require new understandings of how the institutional environment interacts with boundedly rational behavior.

Intrinsic Motivation: What Can Be Achieved Without a Reputation System?

Standard economic theory, based on a narrow definition of self-interest, implies that without external control and incentives, there is hardly any hope that trust and trustworthiness can emerge, but also that a perfect reputation system can create enough incentives to solve the problems. Our work suggests that economic theory is misleading on both counts. There can be trust without external enforcement, and there can be cooperation failure even with perfect reputation systems. Thus, when we attempt to measure the impact of the introduction of a perfect reputation system in a community of strangers, we need to carefully measure both how well the community does absent any external cooperation enforcement and how well it does with a perfect enforcement system. Although in reality neither environment exists, we can create both situations in the laboratory. For instance, we can create situations that are anonymous and truly one-shot for our subjects in the sense that none of the encounters are linked by flows of reputation information.

To make things simple and to abstract away from various complicating factors, we focus on a simple buyer-seller game featuring a trust problem typical of those that reputation systems are commonly used to mitigate. Figure 1.1 illustrates the moves in the buyer-seller encounter. Both the seller and the buyer are endowed with $35, which is the payoff when no trade takes place. The seller offers an item for sale at a price of $35 that has a value of $50 to the buyer. The seller's cost of providing the item is $20. If the buyer chooses to buy the item, he sends the seller his endowment of $35. The seller then has to decide whether to ship the item, or whether to keep both the money and the item. If the seller does not ship, he receives the price plus his endowment of $35 for a total of $70. If he ships, he receives the price minus the costs plus his endowment for a total of $50. If the buyer chooses not to buy the item, no trade occurs.

Figure 1.1 The Buyer-Seller Encounter



Source: Reprinted with permission from Bolton, Katok, and Ockenfels (2004a). Copyright 2004, the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 300, Hanover, Md. 21076, U.S.A.

At the heart of the game is a moral hazard problem that must be overcome if trades are to be successfully executed. With no common history or common future among traders that could give them the opportunity to reward or punish each other, and with no other kind of external (say, legal) enforcement, the seller can profit from not sending the item or sending poorer quality than promised. That is, the seller's pecuniary motive in the figure 1.1 game dictates that he keep the money along with his endowment. In this case, the buyer would lose his endowment and end up with nothing. Anticipating this moral hazard, buyers may not be willing to buy. As a consequence, trading that would make everyone better off would not take place. This is the essential trust dilemma that economic and social interactions—whether they be online or offline—need navigate.[6]

Economic theory presumes that under the given circumstance all rational sellers will fall to moral hazard, and consequently, all trustworthiness, and therefore trust, will vanish. However, the standard models assume that people are guided solely by pecuniary concerns. In reality, people care about other things as well. In fact, in trust games and related anonymous one-shot games (like the prisoner's dilemma game and the ultimatum game), psychologists, sociologists, experimental economists and others have identified several nonpecuniary motives that are important drivers of behavior in these situations. Most prominent in the recent economics literature are concerns for fairness (Fehr and Schmidt 1999; Bolton and Ockenfels 2000) and reciprocity (Rabin 1993; Dufwenberg and Kirchsteiger 2004). These social preference models assume that

traders care about their monetary payoff but that some may also be concerned with the social impact of their behavior. Reciprocity models conjecture that people tend to be kind in response to kindness and unkind in response to unkindness, whereas fairness models posit that some individuals may have a preference for equitably sharing the efficiency gains from trade.[7]

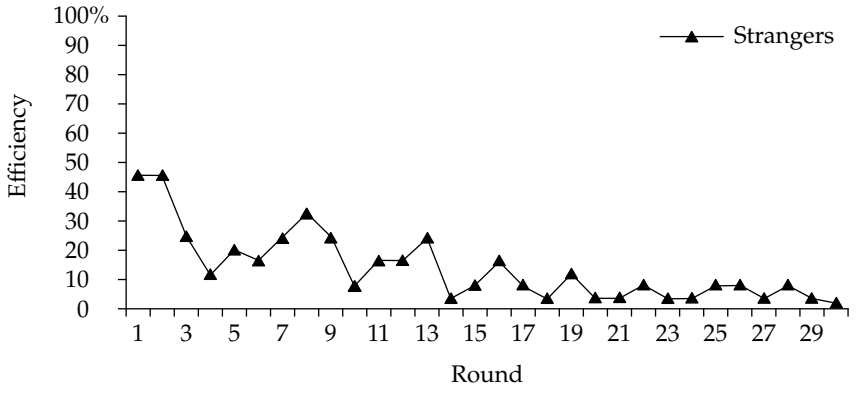
We studied the game in figure 1.1 in a classroom experiment (Bolton, Katok, and Ockenfels 2004b). We found that 37 percent of the thirty sellers were willing to ship in anonymous one-shot encounters and that 27 percent of the buyers were willing to buy. Contrary to the predictions of standard theory, then, there is a nontrivial amount of trust and trustworthiness in anonymous one-shot encounters.[8] At the same time, room for improvement is substantial. On average, only about 10 percent of all encounters ($= 0.27 \times 0.37$) end up in successful and efficient trade. Furthermore, this figure probably overestimates the power of intrinsic motivations to behave reciprocally in a dynamic setting. That is, in expected monetary terms, the probability of a trustworthy seller needs to be at least 70 percent to make buying in the trust game profitable. In our one-shot game, the probability was well below this threshold.

A natural hypothesis, then, is that if trust rests solely on behavioral propensities, trust will diminish over time. This hypothesis has been tested (Bolton, Katok, and Ockenfels 2004a). In our laboratory experiment, the market transactions take place over a series of thirty rounds. At the beginning of each round, a potential buyer is matched with a potential seller and they then play the trust game in figure 1.1. Each game is played with a different transaction partner and no information about trade outcomes leaks from one encounter to another one, so we call this experimental treatment the *strangers market*. All interaction was computer mediated and anonymous; subjects sat in cubicles in front of computers not knowing the true identity of their trading partners, capturing an important aspect of online trading. The rules, and that all rounds would be paid, were common knowledge. Observe that, absent reputation information, this market is essentially a sequence of one-shot games. Thus, because there is not enough intrinsic trustworthiness to make trust profitable in the nonrepeated one-shot game, we hypothesize that buyers quickly learn that cooperation does not pay out and that, subsequently, trading activities will collapse.

Figure 1.2 shows the average buying and shipping (conditioned on buying) behavior across rounds.

Aggregating over all rounds, trustworthiness is about the same as in the one-shot version of the trust game in figure 1.1. This reflects that the strangers market does not create additional incentives to be trustworthy compared to the one-shot game. On the buyer side, there is, on average, more trust in the strangers market than in the one-shot version of the

Figure 1.2 Strangers Treatment



Source: Reprinted with permission from Bolton, Katok, and Ockenfels (2004a). Copyright 2004, the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 300, Hanover, Md. 21076, U.S.A.

game, possibly reflecting the hope that repeated action will support more cooperation. But the dynamics reveal that buyers respond to the fact that, on average, this expectation was disappointed: they start out by trusting quite a lot, but trust quickly collapses. In fact, the percentage of last round trust was only 0.04 percent, much less than in the one-shot game, indicating that buying in the one-shot game is mainly due to inexperience.

In sum, economic theory underestimates the degree of cooperation in one-shot encounters of anonymous traders; there is intrinsic trustworthiness. To the extent people cooperate, the need for a reputation system is diminished. However, in our setting there is not enough intrinsic motivation to stabilize positive reciprocity in an anonymous community without external enforcement. In this sense, economic theory is right: relying on solely intrinsic motivation will not, in the long run, lead to satisfactory cooperative behavior.

Extrinsic Motivation: What Is the Gain of Introducing a Perfect Reputation System?

Here we look at how well reputation systems provide an external enforcement device that may help overcome the cooperation problems in anonymous communities. A number of other external factors influence trust and trustworthiness. Elsewhere in this volume, Karen Cook and

her colleagues discuss how, absent reputation information, competence and motivation can influence trust and trustworthiness (see chapter 7). Tapan Khopar and Paul Resnick discuss the influence of culture (see chapter 4).

From an economic theory perspective, the incentives created by reputation systems depend on the exact trading environment. Suppose, for the moment, that the buyer-seller encounter in figure 1.1 is played repeatedly, with an infinite time horizon, and so with no expectation of a stopping round of play. In such a setting, even if all traders are selfish and rational, equilibria exist in which the buyer always buys and the seller always ships. The equilibria can be supported by reciprocal trigger-strategies that call for a buyer, for instance, to trust as long as the seller has shipped when he or she has had opportunities to do so in the past. Once the seller defects, he will never be trusted again. If future payoffs are important enough, the seller has an incentive to be trustworthy all the time, and the buyer has an incentive to trust all the time (for example, Kandori 1992; Greif 1994). An interesting feature of this argument is that it is independent of whether the reputation system relies on direct or indirect reciprocity. The information available to the buyer about the seller is what is important; if the information is sufficient in quantity and accuracy, the buyer can act on it just as well if the information were generated elsewhere or if it were generated from the buyer's experience.

There are, however, in our context two problems with this kind of simple equilibrium. First, the trading horizon in online market platforms is typically finite. If either the buyer or the seller believes that there will be some upper limit to the number of items to be traded (so a finite horizon game), cooperation among selfish, rational traders will unravel (in the last round there is no trustworthiness, and so no trust and no trade, and for this reason no trade in the second to last round, and so on). Second, buying and shipping in the infinite game equilibrium does not capture trust and trustworthiness under conditions of uncertainty. Specifically, in the infinite game equilibrium, there is no uncertainty about each other's behavior because, in equilibrium, all sellers—not just some or most—have a material incentive to be trustworthy and ship.[9] In this sense, there is no risk of being exploited. Yet, in many cases, in real-world markets a buyer trusts in the sense that they decide to purchase knowing there is some chance of exploitation.

Because this chapter is concerned with trust (characterized by a risk of being exploited) in economic transactions (where traders typically trade a finite number of items), we think it more appropriate to study finitely repeated games. In models of these games, trust emerges when there is some, possibly small, amount of truly intrinsic trustworthiness within the seller population (Wilson 1985). That there is intrinsic trustworthiness has been demonstrated, for instance, in our experimental

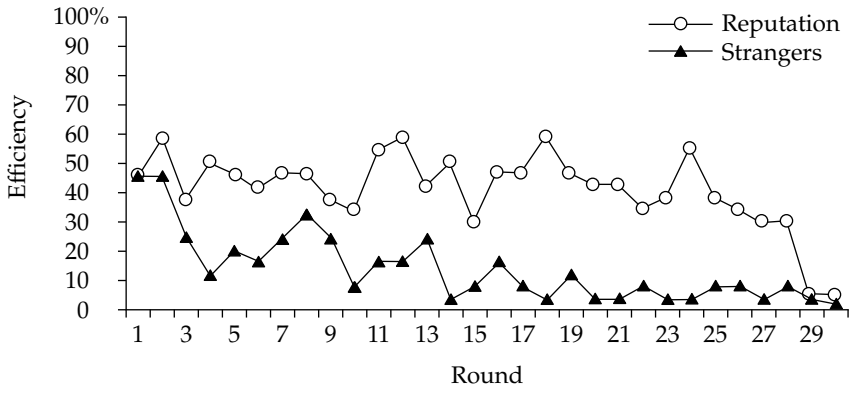
studies of the one-shot trust game of figure 1.1. In essence, in theory, the existence of some intrinsically trustworthy sellers gives all sellers an incentive to build a reputation as trustworthy, at least until toward the end of the game, at which point a good reputation is less valuable. Hence buyers can trust sellers, at least early on, because there is a high probability—albeit less than one in the last few rounds—that all sellers will act trustworthy. [10] It turns out that reputation building in this model, in the context of the buyer-seller encounter, is, as in the infinite horizon models, independent of whether the reputation system relies on direct or indirect reciprocity (Bolton and Ockenfels 2008), something we will come back to shortly.

Economic theory therefore suggests that, in principle, reputation mechanisms of the sort we describe in the introduction can solve many of the trust problems associated with economic transactions. All the various models, finite and infinite horizon alike, suggest that reliable information about past behavior is a necessary ingredient to the emergence of trust, because it allows buyers to avoid sellers who are known as fraudulent and to buy only from sellers who have proved trustworthy in the past. Conditioning trust on the seller's history creates incentives for sellers to build up a reputation for being trustworthy, at least when the end of the market is not too close and maintaining a good reputation is still valuable. A reputation of being trustworthy can be developed and sustained even by completely rational and selfish sellers—as long as the probability of being matched with intrinsically trustworthy sellers is strictly positive. We know from our experimental strangers market that intrinsic trustworthiness alone is not enough to sustain a trading platform that has no reputation system. So does a feedback system help promote trust and trustworthiness, as suggested by theory?

Gary Bolton, Elena Katok, and Axel Ockenfels compared the strangers market to a *reputation market*, played more than thirty rounds, in which, as before, a buyer never met the same seller more than once (2004a). However, in this market we introduced a reputation system that, similar to eBay's feedback forum, informs buyers about all past actions of their current seller (for related experimental work, see Duffy and Ochs 2003; Bohnet and Huck 2004). This feedback information is always shared and reliable, because it is not given by the buyers but by the experimenter, and sellers had no way to change their online identity. This way, the experiment studies the impact of feedback information on trading behavior when an ideal, frictionless reputation mechanism is available. And in the finite horizon theory, this should be enough information to enable trust and trustworthiness, and so successful trade.

Figure 1.3 shows the results of the reputation market experiment and compares with the strangers market results from figure 1.2. On average, there is significantly more buying (56 versus 37 percent; $p < .05$) and shipping (73 versus 39 percent; $p < .01$) in the reputation market than in

Figure 1.3 Reputation Market



Source: Reprinted with permission from Bolton, Katok, and Ockenfels (2004a). Copyright 2004, the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 300, Hanover, Md. 21076, U.S.A.

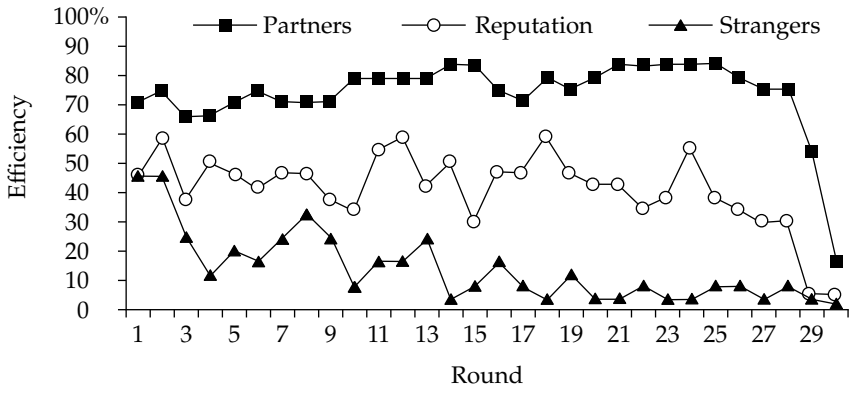
the strangers market. In fact, the shipping probability is slightly higher than the threshold of 70 percent for trust being profitable. As a consequence, the trade dynamics also look quite different than in the strangers market; trading starts at about the same level as in the strangers market and remains stable until the very last rounds, when the strategic value of having a reputation for being trustworthy vanishes and virtually all cooperation collapses.

We conclude that introducing a perfect reputation system in a market with strangers has a strongly positive impact on trust, trustworthiness, and trading efficiency. Both buyers and sellers respond strategically to the information provided. At the same time, however, the experiment demonstrates the serious limits of perfect reputation systems in promoting cooperation. The realized surplus as a proportion of potential surplus is only 41 percent. The gain from introducing a perfect system into a strangers market, described earlier, as a proportion of the maximal potential gain is 41 – 14 = 27 percent, well below what would be expected theoretically (see Bolton and Ockenfels 2008). Obviously, trader behavior is different from what we expect from theory, in a way that limits the effectiveness of reputation systems.

What Behavior Limits the Effectiveness of Reputation Systems?

We have seen that even though reputation systems can build on intrinsic motivations to cooperate, their effectiveness is less than what can be ex-

Figure 1.4 Partners Market



Source: Reprinted with permission from Bolton, Katok, and Ockenfels (2004a). Copyright 2004, the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 300, Hanover, Md. 21076, U.S.A.

pected from theory based on purely selfish traders. What is the source of these limitations?

Because the feedback system in the experiment is perfect with regard to the information it delivers, we need to look at departures from fully rational behavior for answers. Evidence in the data indicates that forward looking behavior is more limited than theory anticipates. Perhaps the strongest evidence for this is that out-of-equilibrium behavior is observed in the early rounds of play (see figure 1.4). For instance, the sellers' payoffs are strongly positively correlated with the overall number of shippings; the Spearman rank correlation is 0.504 ($p = .000$). Shipping early is not only trustworthy and fair, it also pays. However, many sellers have difficulty understanding the future benefits of being nice. About 40 percent of the sellers in the reputation market who receive an order in the first round of the market fail to ship.

Evidence also indicates that traders learn from looking back, a kind of learning the equilibrium model does not anticipate. For example, reputation market sellers are actually more inclined to ship in the middle rounds of the market than at the beginning. And in the strangers market, the 65 percent of buyers who start out trusting quickly learn that they should not. This behavior is consistent with low-rationality adaptive learning models that suggest that people come to strategic games with rough priors and adjust these priors according to the payoff reinforcement they get from experimenting with various strategies (see Erev and Roth 1998). Davide Barrera and Vincent Buskens, in the following

chapter, present data that suggest that another form of learning by looking back, learning by imitation, is also important in games involving trust.

There is also additional evidence to suggest that bounded rationality is not the entire story for why reputation market trading performance falls short of what theory leads us to expect. Bolton and his colleagues also included a *partners market* (2004a). The only respect in which this market differed from the reputation market is that, in the partners market, the same buyer was matched with the same seller for the entire market. Recall that theory suggests that there should be no difference in the performance of the two markets: in both cases, buyers should be able to play tit-for-tat strategies to keep sellers trustworthy. Nevertheless, figure 1.4 demonstrates a substantial difference between them. Overall, trading (efficiency) levels in the partners markets, 74 percent, is significantly higher than in the reputation market ($p < 0.025$).

The amount of trading in the partners markets is still substantially less than perfect, indicating that bounded rationality explanations still apply. Still, trade efficiency is greater than in the reputation markets, which suggests that some other things beside bounded rationality are at play. We argue that the flow of information in the reputation markets creates information externalities in that, out of equilibrium, the incentives to invest in the two markets are different. Specifically, there is a public goods problem in the reputation market not present in the partners market. Buyers do not benefit from the reputation information they themselves produce. As a consequence, reputation market buyers underinvest in the production of reputation information relative to partners markets.[11] In this way, trust is an attribute of the system, not just the individuals in it (for a demonstration of this point in a different market context, see chapter 3, this volume).

So boundedly rational trading is off the equilibrium path, and the resulting out-of-equilibrium incentives may in turn affect traders' behavior. A second observation in this regard is that reputation information, even in a system with comprehensive and reliable feedback information, need be interpreted as a noisy signal because the predictive value of reputation information suffers from the noise generated by the behavior of real traders. This has consequences in a number of ways (for evidence on the relationship between noisy signals in reputation and perceptions of fraud, see chapter 8, this volume). Also, we have experimentally shown in a recent paper that market competition tends to increase the effectiveness of reputation systems in environments with noisy behaviors (Bolton, Loebbecke, and Ockenfels 2008). It does so because, with competition, buyers can discriminate between sellers on the basis of the reputation information provided by the reputation system, creating stronger incentives for sellers to behave consistently trustworthy over

time. The experiments involved matching competition (each buyer gets to choose between two sellers and prices are fixed) and price competition (the two sellers compete on prices) to both the reputation and partners markets described earlier. Our experiments showed that seller competition in (perfect) reputation markets typically enhances trust and trustworthiness, and always increases total gains from trade. We also found that information about reputation trumps pricing in the sense that traders usually do not conduct business with someone having a bad reputation—not even for a substantial price discount. (Andreas Diekmann, Ben Jann, and David Wyder, in chapter 5 of this volume, find that buyers are willing to pay a higher price to sellers with good reputations.) Price competition thus does not significantly undermine the sellers' incentives to be trustworthy. Finally, we found that a reliable reputation system can largely reduce the advantage of partners markets over reputation markets in promoting trust and trustworthiness described earlier, if the market is competitive enough. One important overall conclusion from the study, then, is that in a world with noisy traders and well-functioning reputation systems, encouraging greater market competition may be a powerful tool for increasing cooperation and trade efficiency (Bolton, Loebbecke, and Ockenfels 2008).

Complex Reputation Measures

We have discussed markets where reputation is equivalent with (perfect) information about the sequence of a seller's shipping decisions. In theory, this measure is enough to sustain cooperation through indirect reciprocity. In fact, simple and stable cooperation in these market settings can theoretically be reached with just information about a seller's last shipping decision, because this information is all one needs to use tit-for-tat strategies. Here we look at markets that are more complex and, in theory, require more reputation information to produce cooperation.

We examine two types of complexities. One arises in markets where reputation must be built on multidimensional facets of the seller's history. When assessing a seller's trustworthiness, buyers need to take into account, for instance, technical and cultural communication problems, the possibility of incomplete or manipulative feedback, the reliability of the postal service, and so on. Even perfect reputation mechanisms, which deliver all relevant information to promote cooperative interaction, may become quite complex, so that real traders experience information overload.

A second type of complexity arises from information requirements in two-sided reputation systems, which are necessarily much more demanding than one-sided systems. For example, consider a system in

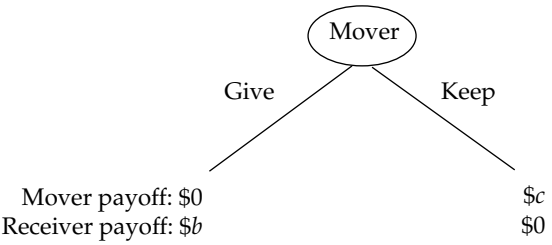
which buyers rate sellers and sellers rate buyers to mitigate moral hazard incentives on both market sides.[12] Now suppose that a buyer receives reliable information that the seller did not send the object to the last buyer. Does this imply that the seller is not trustworthy? No. It could be that the seller did not ship because his or her last buyer never sent the payment. Let's think this one step further. Would it then be enough for our current buyer to know whether the current seller's last buyer paid? Again, the answer is no. Whether the last buyer's action can be interpreted as trustworthy depends on the history of play of his or her earlier transaction partners. In principle, the entire history of both trading partners as well as their trading partners, and their trading partners, and so on, may be required to construct a system that has enough information of the sort we tested in the one-way settings discussed earlier.

Clearly, this information is difficult to process, even when comprehensive and perfectly reliable. There is a way to avoid the processing problems, though. The relevant information can be captured in a single reputation rating, which does not directly reveal past behavior but rather evaluates these behaviors according to all traders' histories and with respect to a trading norm. This rating can, in theory, be easily processed. On the other hand, however, the information content is less comprehensible because of the rather complex information aggregation processes behind the rating.

Let us illustrate the issues with the help of the simple image scoring game (Nowak and Sigmund 1998). As with the markets we have already studied, the image scoring game conceives of the group interacting over a series of rounds. Again, in each round, people are paired off at random. One person in the pair, designated as the *mover*, is given the opportunity to give a favor to the other, designated as the *receiver*. These designations are assigned randomly, so over many rounds, each player is a mover about half the time and a receiver the other half. Giving a favor would cost the mover c and benefit the receiver $b > c > 0$. Figure 1.5 illustrates the situation.

The efficient outcome in this game, the outcome that maximizes the total social benefits, is for everyone to give when they are the mover. Although keeping maximizes short-run payoffs, reputation can help by providing the information necessary to reward those who give with giving and punishing those who do not with keeping. This kind of reciprocity is not unlike the trust game context we discussed in earlier sections. However, even though the game looks much simpler than the one presented in figure 1.1, the basic reputation issue is more complicated. To see why, consider the kind of reciprocity that works in the trust game markets based on the game in figure 1.1. The mover gives if he knows the receiver played give the last time as a mover, and keeps if the receiver last played keep. Suppose now that you are the mover matched

Figure 1.5 Mover Meets Receiver in Image Scoring Game



Source: Authors' compilation.

with someone who last played *keep* as a mover. If you play keep as the reciprocity strategy stipulates, then the next time you are the receiver, you can expect the mover to play keep on you (if others too play the reciprocity strategy). Consequently, you make more money playing *give* (lose c now, pays b later) than playing keep (gain c now, pays 0 later). The problem is that if enough people decide to give to keepers then it pays to be a keeper. And if it pays to have a bad reputation, then why have a good one?

So, this kind of first-order information about what the opponent did last time as a mover is, theoretically, not enough to stabilize cooperation. If we add second-order information, the receiver's reputation would include not only what he did last time as a mover, but also what the receiver he faced did last time as a mover. For example, the reputation might reveal that the receiver last played keep with a player who last played give. This amount of recursive information pushes the unraveling problem back by a step. To see this, consider a mover who, for the first time, encounters a receiver who played keep on a giver. To support his punishment, keeping on a keeper would have to be rewarded, meaning that there needs to be giving to someone who gives to a keeper—which is not consistent with self-interest because keeping on a keeper pays more. So now players would have to think two steps ahead and be confident others do so as well before cooperation would unravel.

Of course, thinking three steps ahead is still not enough. To stabilize cooperation in a population of rational traders, one would need the entire transaction histories of basically all traders. For this reason, some theorists have cautioned that indirect reciprocal systems might not be stable outside very small groups, where information demands are relatively modest.

Boundedly rational traders, however, often do not think many steps

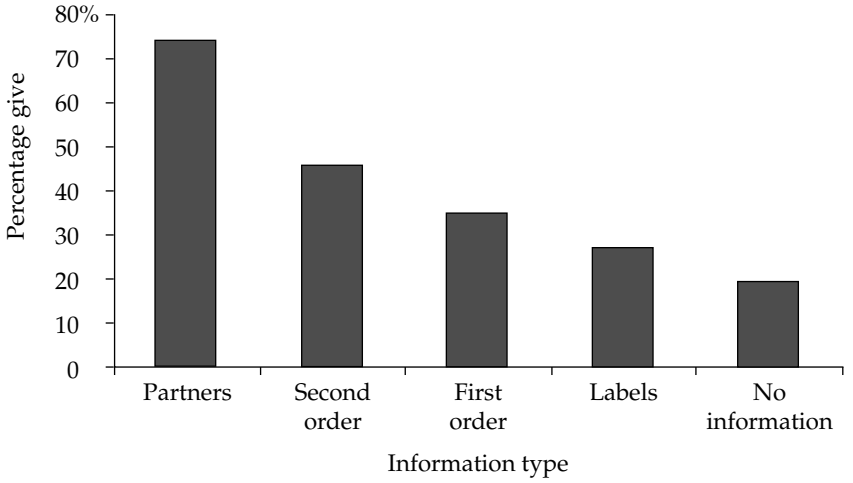
ahead (see earlier and, for example, Nagel 1995), and people's ability to do backwards induction is rather limited. In fact, in an experimental study of the image scoring game, we find that first-order information significantly increases cooperation rates above the level in a market without any reputation information (for details, see Bolton, Katok, and Ockenfels 2005).[13] Second-order information again significantly increases cooperation rates, reflecting that traders do some of the backward induction, but do not think through the whole problem. However, both markets with strangers matching perform dramatically worse than the corresponding partners market. Figure 1.6 illustrates the situation.

How can the gap of the effectiveness of reputation systems between partners and strangers matching markets be closed? We think it unlikely that higher-order information would help considerably, because second-order information is already difficult to communicate and to process. One way could be to aggregate all the relevant information into a single reputation score so that traders might then apply a simple reciprocity strategy in a way that cannot be cheated on (Kandori 1992).

In our experiment (Bolton, Katok, and Ockenfels 2005), we proposed the following reputation score along the following lines. We labeled each player in each round as a member of either the *matcher* club or the *nonmatcher* club according to the following rules. In the first round, everyone is a matcher. In every round after that, a player's label is updated: If the player gave to a matcher the last time he was mover, he is a matcher. If the player kept on a nonatcher, he is a matcher. If the player did anything else, he is a nonmatcher. Now consider a reciprocity strategy that stipulates giving to a matcher and keeping on a nonmatcher. If everyone follows this rule, then everyone will stay a matcher and there will be 100 percent cooperation. Moreover, you cannot benefit by cheating. If you keep on a matcher, you become a nonmatcher, which lines you up to be punished because the next time you are matched with a mover, he will keep on you. And punishment is now with impunity: keeping on a nonmatcher allows a mover to maintain matching status—he won't be punished for doing the right thing.

When all information is processed in this way, the reciprocal strategy yields stable cooperation—at least in theory. To our surprise, however, the experimental data does not confirm at all the prediction. The bar called *labels* in figure 1.6 shows the average giving rate in this setting. The information that should stabilize the cooperation rate in fact significantly reduces the cooperation rate compared to the other settings, which involve theoretically insufficient reputation information. It appears that real traders have difficulties with reputational reports that filter actions, and respond more favorably to reputational reports about recent past actions. The dilemma is that this information, when complete, cannot be processed by boundedly rational traders.

Figure 1.6 Giving Levels (Averaged over All Rounds)



Source: Authors' compilation.

Conclusions

What we learn from the experimental and theoretical work is that it is the interplay of institutions with bounded rational behavior that drives the results. No doubt, institutions matter, but behavioral aspects of reputation-building matter as well. As a result, standard economic models based on full rationality and narrow self-interest tend to overestimate the difficulties of promoting trust in one-shot situations, and underestimate the difficulties in ongoing interaction in communities of strangers.

Because the laboratory reputation systems we study here are perfect, the limits of their effectiveness cannot be the result of institutional defects but must be due to behavioral defects. That is, the restraints that we observe are rooted in boundedly rational behavior. There are basically two types of noisiness in the behavior that significantly affect the functioning of the institutions. For one, bounded rationality can directly affect trust and trustworthiness though nonrational choices. Besides difficulties that arise when handling complex reputation measures, we observe that real traders have difficulties coping with reputation-building dynamics. Many traders fail to look forward enough and to fully take into account the future consequences of current behavior. Other behaviors are characterized by too much backward looking and simple, adap-

tive learning patterns. Second, noisy behavior moves the reputation-building dynamics off the equilibrium path and thus changes (out-of-equilibrium) incentives in ways that systematically affect strategic reputation-building. We observe, for instance, that noisy behavior creates information externalities so that the flow of reputation information through the community becomes critical to the effectiveness of reputation systems. Also, when trading dynamics are out of equilibrium, seller competition becomes a powerful support for reputation systems.

We think that only a combination of complementary field, laboratory, and thought experiments can reveal the full story behind reputation systems. Field studies strive for external validity and require a careful look at institutions. It can be difficult, if not impossible, however, to separate institutional from behavioral influences, to measure the impact of either aspect on the effectiveness of reputation systems, and to measure the overall impact of a reputation system (for an in-depth discussion of the limitations of field data and techniques that might be used to overcome them, see chapter 6, this volume). Thought experiments (for example, equilibrium theory) help us understand how behavior and institutions interact, reveal basic incentive structures, and allow generalizing from empirical observations. But it is risky not to complement thought experiments with data because it is known that theory can sometimes yield dramatically wrong conclusions, especially when it comes to social interaction (for example, Bolton and Ockenfels 2000). Thought experiments also tend to neglect institutional details, which can turn out to be critical, both in the equilibrium analysis and in reality (for example, Klemperer 2004). Laboratory experiments can separate and measure the different impacts and the interplay between institutional and behavioral influences. Combined with field and thought experiments, they are a powerful tool for analyzing the effectiveness of existing and newly designed reputation systems.

Gary Bolton gratefully acknowledges the support of the National Science Foundation. Axel Ockenfels gratefully acknowledges the support of the Deutsche Forschungsgemeinschaft. We are advising firms, including eBay, on reputation mechanism design and other market design issues; the views expressed are our own.

Notes

1. In this chapter, we deal with trust and trustworthiness in Internet marketplaces. For a taxonomy of information exchange systems, see chapter 10.
2. Although online auction transactions appear to be particularly vulnerable to fraud, the problems we report here exist in basically all reputation-based interaction. We consider eBay a convenient example because it allows re-

searchers to quantify some of the benefits and problems. We also note that eBay's feedback forum is only one part of a mix of (imperfect) policies and rules that interact to promote trade efficiency. Only a very few papers address this interaction. One is that of Werner Güth and his colleagues, who investigate the joint effectiveness of buyer insurance, which is part of eBay's so-called Purchase Protection Program, and eBay's feedback forum (Güth, Mengel, and Ockenfels 2006).

3. The cost of generating feedback includes the risk of trusting sellers, something we discuss later.
4. One of the main motives for giving feedback appears to be reciprocity (Dellarocas, Fan, and Wood 2004). That is, a trader's propensity to leave feedback is driven by the expectation that the trading partner reciprocates with positive feedback. This observation is remarkably in line with the literature in experimental economics on voluntary provision of public goods (see, for example, Ledyard 1995; Ockenfels and Weimann 1999).
5. Another potential source of noisy feedback information is fraudulent identity change. The costs of changing an online trader identity is often close to zero, implying that fraudulent sellers can exploit their buyers and then reappear with a clean record. If buyers are willing to buy only from a newbie, a seller with no record, if the object is offered at a lower price, compared to the price offered by a seller with a positive reputation record, then trust and trustworthy behavior can be sustained (Ockenfels 2003; see also Friedman and Resnick 2001).
6. We assume that the seller fixes the price. For example, Amazon.com permits sellers of used books and CDs to make offerings on its site, along with its own new goods offerings. A used goods seller posts on the market platform an offer that includes a description of the item and its condition, and a price at which he or she is willing to sell. A willing buyer sends the money to Amazon. On receiving the money, the seller is supposed to ship the item to the buyer. In addition, the moral hazards surrounding shipping and accurate representation of good quality are controlled by a feedback system not unlike the one we will introduce to our game. However, all arguments in this chapter hold equally if the price is endogenously determined, such as in eBay's auctions (in this case the auction winner is the buyer).
7. To be more specific, in our trust game, reciprocity models suggest that a seller ships because the buyer was so kind to buy, whereas fairness models suggest that the seller ships because otherwise the payoff distribution would be unfair (see also the discussion of motives like efficiency concerns and procedural fairness in related games in Bolton and Ockenfels 2008).
8. Payoffs, framing, and context may all affect the exact numbers. However, based on our extensive research with various payoff parameters, different framings, contexts, and experimental procedures (see reference list), we are confident that the qualitative results we discuss in this chapter are robust. One element we do not consider is buyer-seller verbal communication. For data suggesting that such communication can mitigate moral hazard, see chapter 11, this volume.
9. There are also more subtle equilibria in these models in which cooperation in any given round is uncertain, but this raises yet a third problem—that

there are many equilibria in these models with outcomes ranging from full cooperation to no cooperation. In our view, trust is not satisfactorily described as an equilibrium selection problem.

10. The mechanics of these equilibria are relatively complex, and we will not delve into them here (for a theoretical and experimental treatment within a trust game environment, see Bolton and Ockenfels 2008).
11. It turns out that there are no information externalities in the incomplete information model of reputation building on the equilibrium path of the buyer-seller encounter or in other market games such as Selten's chain store game (for a discussion and a formal experiment that shows that the phenomenon is more robust than theory suggests it should be, see Bolton and Ockenfels 2008).
12. This is much like eBay, where both transaction partners can rate each other. However, because eBay transactions are typically sequential (first buyers send money, then sellers ship the object), the moral hazard problem is mostly on the seller side.
13. Subjects were Penn State University students, mostly undergraduates from various fields of study, and were recruited by fliers posted around campus—in total, 192 participants. We ran two image scoring games for each information condition, each game with sixteen subjects playing for fourteen rounds. In each round, subjects were anonymously paired, interfacing with one another by computers. The value of a gift, B, was $1.25, and the cost of giving, C, was $0.75. Subjects knew that they would be in each role, mover or receiver, for half the trials (seven times), and roles would generally rotate between rounds.

References

Bohnet, Ing, and Steffen, Huck. 2004. "Repetition and Reputation: Implications for Trust and Trustworthiness When Institutions Change." *American Economic Review* 94(2): 362–66.

Bolton, Gary E., and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity and Competition." *American Economic Review* 90(1): 166–93.

———. 2008. "Information Value and Externalities in Reputation Building: An Experimental Study." Working paper. Cologne: University of Cologne.

Bolton, Gary E., Ben Greiner, and Axel Ockenfels. 2008. "Engineering Trust: Reciprocity in the Production of Reputation Information." Working paper no. 42. Cologne: University of Cologne.

Bolton, Gary E., Elena Katok, and Axel Ockenfels. 2004a. "How Effective Are Online Reputation Mechanisms? An Experimental Investigation." *Management Science* 50(11): 1587–602.

———. 2004b. "Trust among Internet Traders: A Behavioral Economics Approach." *Analyse & Kritik* 26(2): 185–202.

———. 2005. "Cooperation among Strangers with Limited Information about Reputation." *Journal of Public Economics* 89(8): 1457–68.

Bolton, Gary E., Claudia Loebbecke, and Axel Ockenfels. 2008. "How Social Reputation Networks Interact with Competition in Anonymous Online Trad-

ing: An Experimental Study." *Journal of Management Information Systems* 25(2): 145–69.

Dellarocas, Chrysanthos. 2003. "The Digitization of Word-of-Mouth: Promise and Challenges of Online Reputation Mechanisms." *Management Science* 49(10): 1407–424.

———. 2006. "Reputation Mechanisms." In *Handbook on Economics and Information Systems*, edited by Terrence Hendershott. Amsterdam: Elsevier Science.

Dellarocas, Chrysanthos, Ming Fan, and Charles Wood. 2004. "Self-Interest, Reciprocity, and Participation in Online Reputation Systems." MIT Sloan School of Management working paper no. 4500-04. Cambridge, Mass.: Massachusetts Institute of Technology.

Duffy, John, and Jack Ochs. 2003. "Cooperative Behavior and the Frequency of Social Interaction." Working paper no. 274. Pittsburgh: University of Pittsburgh.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior* 47(2): 268–98.

Erev, Ido, and Alvin E. Roth. 1998. "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria." *American Economic Review* 88(4): 848–81.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114(4): 817–68.

Friedman, Eric J., and Paul Resnick 2001. "The Social Cost of Cheap Pseudonyms." *Journal of Economics and Management Strategy* 10(2): 173–99.

Greif, Avner 1993. "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition." *American Economic Review* 83(3): 525–48.

———. 1994. "Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies." *Journal of Political Economy* 102(5): 912–50.

Güth, Werner, Friederike Mengel, and Axel Ockenfels. 2007. "An Evolutionary Analysis of Buyer Insurance and Seller Reputation in Online Markets." *Theory and Decision* 63(3): 265–82.

Kandori, Michihiro. 1992. "Social Norms and Community Enforcement." *Review of Economic Studies* 59(1): 63–80.

Klein, Tobias J., Christian Lambertz, Giancarlo Spagnolo, and Konrad O. Stahl. 2007. "Reputation Building in Anonymous Markets: Evidence from eBay." Working paper. Mannheim: University of Mannheim.

Klemperer, Paul. 2004. *Auctions: Theory and Practice*. Princeton, N.J.: Princeton University Press.

Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research." In *Handbook of Experimental Economics*, edited by John H. Kagel and Alvin E. Roth. Princeton, N.J.: Princeton University Press.

Miller, Nolan, Paul Resnick, and Richard Zeckhauser. 2005. "Eliciting Honest Feedback: The Peer Prediction Method." *Management Science* 51(9): 1359–73.

Nagel, Rosemarie. 1995. "Unraveling in Guessing Games: An Experimental Study." *American Economic Review* 85(5): 1313–26.

Nowak, Marin A., and Karl Sigmund. 1998. "Evolution of Indirect Reciprocity by Image Scoring." *Nature* 393(June 11): 573–77.

Ockenfels, Axel. 2003. "Reputationsmechanismen auf Internet-Marktplattformen." *Zeitschrift für Betriebswirtschaft* 73(3): 295–315.

Ockenfels, Axel, and Joachim Weimann. 1999. "Types and Patterns: An Experimental East-West-German Comparison of Cooperation and Solidarity." *Journal of Public Economics* 71(2): 275–87.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83(5): 1281–302.

Rashid, Al Mamunur, Kimberly Ling, Regina D. Tassone, Paul Resnick, Robert Kraut, and John Riedl. 2006. "Motivating Participation by Displaying the Value of Contribution." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2006,* edited by Rebecca Grinter, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries, and Gary Olson. New York: ACM.

Resnick, Paul, and Richard Zeckhauser 2002. "Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System." In *Advances in Applied Microeconomics*, vol. 11, *The Economics of Internet and E-commerce*, edited by Michael R. Baye. Amsterdam: Elsevier Science.

Resnick, Paul, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. "The Value of Reputation on eBay: A Controlled Experiment." *Experimental Economics* 9(2): 79–101.

Wilson, Robert B. 1985. "Reputations in Games and Markets." In *Game-Theoretic Models of Bargaining*, edited by Alvin E. Roth. Cambridge: Cambridge University Press.