The Evolution of Trust and Reputation: Results from Simulation Experiments

Workingpaper, 21.07.2005

Andreas Diekmann, Department of Social Sciences and Humanities, Swiss Federal Institute of Technology Zurich, ETH-Zentrum, SEW E 26, CH-8092 Zurich, diekmann@soz.gess.ethz.ch

Wojtek Przepiorka, Department of Social Sciences and Humanities, Swiss Federal Institute of Technology Zurich, ETH-Zentrum, SEW E 22, CH-8092 Zurich, przepiorka@soz.gess.ethz.ch

Abstract

In online interactions in general, but especially in interactions between buyers and sellers on internet-auction platforms, the interacting parties must deal with trust and cooperation problems. Whether a rating system is able to foster trust and cooperation through reputation and without an external enforcer is an open question. We therefore explore through ecological analysis different buyer and seller strategies in terms of their success and their contribution to supporting or impeding trust and cooperation. In our agent-based model, the interaction between a buyer and a seller is defined by a one-shot trust game with a reputation mechanism. In every interaction, a buyer has complete information about a seller's past behavior. We find that cooperation evolves under two conditions even in the absence of an external sanctioning authority. On the one hand, some minimal fraction of buyers must make use of the sellers' reputation in their buying strategies and, on the other hand, trustworthy sellers must be given opportunities to gain a good reputation through their cooperative behavior. Despite the apparent usefulness of the reputation mechanism, a small number of deceitful sellers are able to hold their ground.

Problems of Trust and Cooperation

Internet-auction platforms represent markets in which a multitude of transactions between anonymous buyers and sellers are conducted every day. Due to conditions of asymmetric information, the parties to the interaction must address trust and cooperation problems. Mostly, a seller deals with this problem by insisting on payment in advance, thereby, protecting himself from deceitful buyers. The seller ships the good only after receiving payment from the buyer. The buyer therefore must be confident of the seller's willingness to ship the good. The interaction between the buyer and the seller can be formalized as a trust game which can be described as follows (see Figure 1). First, the buyer decides whether to trust the seller or not. If the buyer decides not to trust the seller, the interaction terminates and both agents receive punishment (P) for the missed opportunity. If the buyer decides to trust the seller, a transaction takes place. In this transaction, the buyer transfers a valuable commodity to the seller (e.g. money). Then the seller decides whether to honor the trust shown by the buyer and to repay it with a commodity of equivalent value (e.g. a mobile phone) or to abuse the buyer's trust and to keep the money. If the seller decides to honor the trust shown by the buyer, both agents receive reward (R) for mutual cooperation. If the seller deceives the buyer, the seller receives the gain from temptation (T) while the buyer loses with a sucker's payoff (S) (Dasgupta 2000, Coleman 1990, Buskens and Raub 2002, Bolton, Katok and Ockenfels 2004).

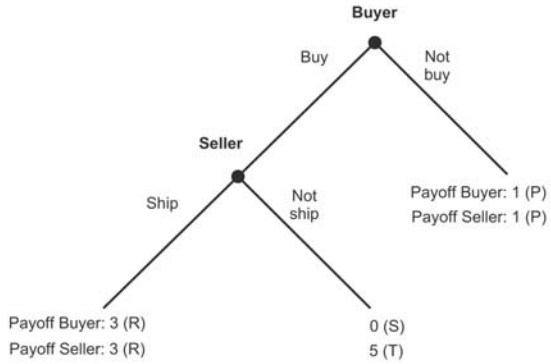

Figure 1: Trust Game

From a rational-choice perspective, neither the seller nor the buyer in a one-shot trust game would have an incentive to behave cooperatively. A "rational" seller would not ship a good that has already been paid for ($T > R$) and a "rational" buyer would not enter the transaction in anticipation of this outcome ($P > S$). The same logic applies when the buyer is the second mover. One might therefore ask how internet-auction markets can evolve at all and why they are not disrupted by dishonest behavior.

The social-dilemma situation described above is comparable to the well-known Prisoner's Dilemma. For each of two interacting agents, it is more rational to defect while mutual cooperation would mean that both parties were better off. As we have learned from Axelrod's (1984) computer tournaments, this socially inferior outcome can be avoided if the "shadow of the future" is high enough to foster mutual cooperation. An agent's expectation that it will interact with the same partner an indefinite number of times and will have the possibility to both reciprocate cooperation and retaliate

for defection makes cooperation more rewarding than it would be in a one-shot game.
However, these conditions can not be assumed to be given in an actual internet-auctions market. Repeated interactions between the same two parties are rather rare (Resnick and Zeckhauser 2002), as market transactions are mostly processed in one-shot interactions. Furthermore, since market participants are anonymous, the possibility of either reciprocating or retaliating is very limited. The simple institution of a rating system can therefore play a crucial role in allowing the market to function.

Introducing Reputation

Rating transactions and then making these ratings available to all interested actors is an important factor in promoting cooperation. In repeated one-shot dilemma games, reputation is a substitute for the iteration of games involving the same two parties. While there is no possibility to reciprocate the cooperation or to retaliate for the defection of one's partner in previous interactions, reputation does allow for indirect sanctioning or reciprocity respectively (Nowak and Sigmund 1998, Wdekind and Milinski 2000, Leimar and Hammerstein 2001, Bolton, Katok and Ockenfels 2005). However, two assumptions have to be met in order for a reputation system to work in the way it is intended: (1) Past behavior must be a predictor for current behavior and (2) a reputation index must capture past behavior (see also Snijders and Zijdeman 2004).
Under these assumptions, a seller with a good reputation has behaved mostly cooperatively and has been rated accordingly, whereas a seller with a bad reputation has been evaluated negatively for her fraudulent past behavior. A buyer having to decide whether to trust a seller or not is assumed to prefer a seller with a good reputation, since this seller is more likely to behave cooperatively again. This difference implies that sellers with a bad reputation are discriminated against and sanctioned (indirectly) for their past behavior. Analogously, sellers with a good reputation are rewarded.
Buyers are thus able to protect themselves from exploitation by considering a seller's reputation before conducting a transaction with that seller. If buyers trust more in suppliers with a good reputation, prefer to trade with them and are perhaps even willing to pay a higher price for transactions with them (premium for reputation), sellers have an incentive to invest in their reputations through cooperation. Sellers on the other hand can insist on payment in advance in order to rule out opportunistic behavior on the part of the buyers (Diekmann and Wyder 2002).

Motivation for Simulation Experiments

Empirical data from internet auctions support the above hypothesis of the price premium on reputation, suggesting that buyers are willing to pay an "insurance fee" to reputable sellers in order to reduce the risk of being deceived (Diekmann and Wyder 2002, Snijders and Zijdeman 2004, Berger and Schmitt 2005). But there are objections to the idea that the reputation system alone ensures the smooth functioning of the internet-auctions market. It has been argued that other protection, specifically the threat of external sanctions by the platform operator, is also necessary in order for cooperative interactions to evolve (Brinkmann and Meifert 2003). Whether cooperation can evolve endogenously in the presence of only a reputation system and in the absence of an external enforcer is therefore a matter of debate. Hence, this question can not be answered through observations of real auctions on existing internet platforms because the threat of sanctions can neither be observed directly nor manipulated freely outside of an experimental setting. We therefore research this question with simulation experiments that reproduce the essential characteristics of internet-auctions markets. In a first passat researching this question, we start from the trust game with a simple reputation mechanism (Dasgupta 2000, Lahno 1995).
In our simulation experiments, the assumption that a reputation index measures past behavior is entirely met. A buyer-agent rates a seller-agent positively, if the trust given by the buyer has been rewarded and rates the seller-agent negatively, if trust has been abused. In every interaction with a seller, a buyer knows the number of negative and positive ratings the seller has been given by other buyers. Although past behavior of an agent might be a good predictor for the agent's behavior in the current interactions the same information on past behavior can be implemented in many ways. We therefore explore whether there are (1) seller and buyer strategies that are more successful than other strategies and (2) successful strategies that are able to promote or impede the emergence of trust and cooperation among the trading population in general. Our aim is to answer the question of whether trust and cooperation can evolve in a market consisting of isolated and anonymous participants solely through a reputation system and without an external enforcer. Our focus lies less on the reputation system itself than on the conditions under which trust and cooperation evolve in the absence of an enforcer.

Simulation Experiments

In a population consisting of n agents, buyers and sellers interact with each other. The buyer population is the same size as the seller population ($m = n/2$). In one round, the two populations interact through every buyer being randomly matched with a seller. Hence, m interactions take place in one round. The interaction between a buyer and a seller is defined by a trust game (see Figure 1 and the description above) and the payoffs are set to $T=5$, $R=3$, $P=1$, $S=0$.
If a transaction between a buyer and a seller has taken place, the buyer must rate the seller in the following

way. If the trust shown by the buyer has been abused by the seller, the seller receives a negative rating ($b = -1$). If the trust shown by the buyer has been honored by the seller, the seller receives a positive rating ($b = 1$). Note that the buyer is never rated by the seller and a seller to whom trust was refused is not rated either.

A generation consists of a specified number of rounds. In every round an agent interacts with a randomly assigned opponent. Given those parameters, it is improbable that the same buyer-seller pair will occur twice in a generation. Repeated interactions between the same two agents are therefore not accounted for. In every interaction, the buyers decide on the basis of an individually defined strategy whether to trust the sellers or not. In every transaction that takes place as a result of the trust shown by the buyers, the sellers decide on the basis of their own individually defined strategies whether to honor the buyer's trust or not.

Sellers only know their own interaction history. Thus sellers do not know how often a buyer showed no trust. They do, however, know something about the mistrust of buyers in general, since they know how often they were exposed to it themselves. Buyers know their own interaction history and the whole transaction history of the sellers with whom they interacted within the same generation. Explicitly, the buyers know the number of negative (u) and positive (v) ratings a seller has received, as well as his reputation index r. The reputation index r is calculated from his positive ratings as a fraction of the total number of his transactions ($r = v / (u + v)$). A seller who has not yet engaged in a transaction has no ratings and, as a consequence, a reputation index of $r = 0$.

The buyer population and the seller population each consist of $m = 256$ agents. In every generation, every seller and every buyer is involved in k interactions (the length of a generation is therefore k rounds). In order to avoid end-round effects, the generation length is not known before the simulation is executed. At the beginning of the first generation, the buyer strategies are distributed in equal proportions throughout the buyer population; the same applies to the distribution of seller strategies in the seller population. At the end of every generation, the payoffs that all agents with the same strategy have obtained are summed into a total score. Then, in the next generation the strategies are distributed throughout the population of buyer and seller agents respectively, in proportion to the scores they have obtained. Neither history, ratings nor the reputation of a parent agent are transferred to a child agent in the next generation.

Results

First, we conduct simple scenario experiments in which we test different ad-hoc buyer and seller strategies. Therewith, we explore the effects on the level of trust and cooperation in the system the properties of the employed strategies and the combination of those have. Second, we arrange tournaments and ask participants to submit their own strategies to compete against other strategies. The tournament approach allows us both to make use of "distributed natural intelligence" in deriving challenges to the mechanisms of our artificial market and to check our experimental findings.[1]

Scenario 1

In the first scenario experiment we start with a buyer population, half of which consists of agents who always give trust to a seller (ALL C). The other half is composed of agents who never give trust to a seller (ALL D) (see Figure 2a). The seller population in turn is dominated by agents that always honor trust. However, one percent of the seller population consists of agents that never honor trust given by a buyer (see Figure 2b). Although the initial fraction of untrustworthy sellers is very small, these sellers manage to invade the population of trustworthy sellers after 15 generations.

At first, the seller population is trustworthy at a very high rate. Agents playing ALL D in the buyer population die out quite early because giving trust is a more successful strategy. After the distrustful buyers have died out, the abuse of trust starts to spread over the seller population. Figure 2c shows that in the beginning, trust is honored at a very high rate, but disappears from the market after the mistrustful buyers' interactions in which no trust is given. What is left is a market of exploiting sellers and exploitable buyers.

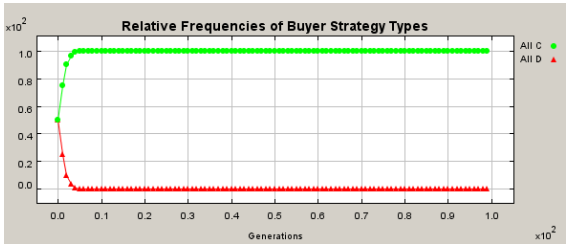

Figure 2a: Fractions of strategy types in the buyer population for 100 generations

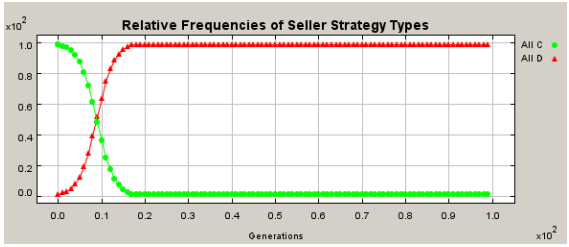

Figure 2b: Fractions of strategy types in the seller population for 100 generations

[1] In the scenario experiments the number of rounds in one generation (k) was 10 and for the tournaments the generation length amounted to k=100 rounds.

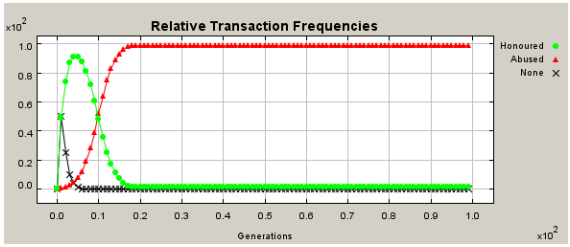

Figure 2c: Fractions of transaction types for 100 generations

Scenario 2

In order to protect trustful buyers from exploitation, we introduce a rating mechanism and a buyer strategy which accounts for a seller's reputation. The initial seller population in scenario 2 consists one half each of trustful and deceitful sellers (see Figure 3b). The buyer population contains, besides the two unconditional strategies ALL C and ALL D, the strategy RT 25. Strategy RT 25 only gives trust to a seller whose reputation index is greater than or equal to $r = 0.25$.[2] The initial fraction of both unconditional strategies is 46% of the buyer population. Accordingly, RT 25 starts at a fraction of 8% of the buyer population (see Figure 3a).

From the beginning, buyers playing RT 25 discriminate against sellers who do not have a positive reputation. This causes a decrease of the fraction of sellers playing ALL D on the one hand and an increasing fraction of buyers playing RT 25 on the other hand. The strategy RT 25 is more successful than ALL C because it does not get exploited by deceitful sellers. As soon as there are enough trustful sellers and a low rate of sellers playing ALL D respectively, buyers playing ALL C become more successful. Because of their unconditional cooperation, buyers playing ALL C also give trust to sellers without a reputation whereas buyers playing RT 25 do not cooperate with sellers who have not yet established a positive reputation. With the rising success of buyers playing ALL C, sellers playing ALL D have more opportunities to abuse trust and in turn become more successful themselves. At the peak of the deceitful sellers' success, again, the fraction of buyers playing RT 25 increases while the fraction of buyers playing ALL C decreases. This coevolutionary dynamics is well mapped by the transaction frequencies depicted in Figure 3c. In terms of the interdependence between the abuse of trust, reputation, mistrust, and trust, the dynamics can be described as follows. Starting with a growing rate of abused trust, the reputation mechanism ensures untrustworthy sellers to be recognized and discriminated against. A rising rate of mistrust due to the distrustful buyers' preference of sellers with a positive reputation is followed by a more confident buyer population and a higher rate of honored trust. The approximate averages of transaction frequencies level off at 60% for Honored Trust, 25% for Mistrust and 15% for Abused Trust.

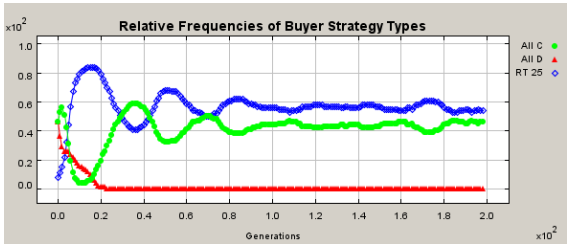

Figure 3a: Fractions of strategy types in the buyer population for 200 generations

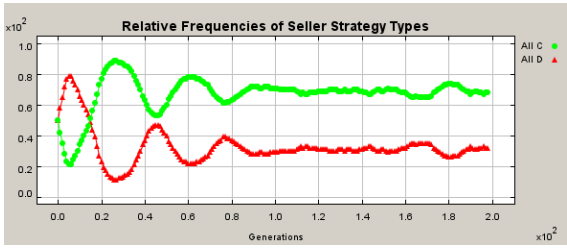

Figure 3b: Fractions of strategy types in the seller population for 200 generations

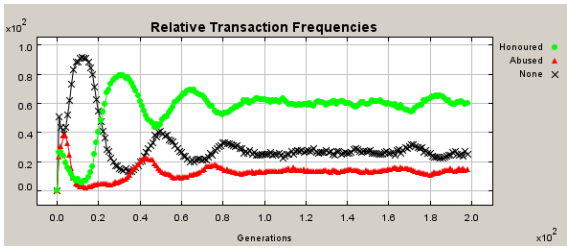

Figure 3c: Fractions of transaction types for 200 generations

However, if the initial fraction of agents playing RT 25 is below 7%, buyers playing ALL C die out too early, depriving trustworthy sellers of the possibility to build a good reputation. Without a reputation, the strategy RT 25 is not able to distinguish between trustworthy and deceitful sellers and therefore would never give trust in any interaction. Without trustful buyers no transactions take place and without transactions a market does not persist.

Scenario 3

The results from scenario 2 suggest that trustworthy sellers must get the opportunity to build a positive reputation. Otherwise, conditional strategies like RT 25 would not be able to distinguish between trustworthy and deceitful sellers. A buyer strategy which incorporates that ability and, at the same time, gives sellers without a reputation the opportunity to prove their trustworthiness, is independent of the existence of a more trustful strategy like ALL C. In this scenario, we introduce the strategy RT 25 C. The only difference between RT 25 and RT 25 C is that RT 25 C unconditionally cooperates in its first interaction. Additionally, the initial fraction of buyer agents playing RT 25 C is 6%. Note that at this initial level, the strategy RT 25 failed to establish and maintain a system based on trust and cooperation.

[2] RT stands for Reputation Threshold. Note, the actual threshold does not matter here since ALL D has always a reputation index of r = 0 and ALL C one of r=1, except in the first round of a generation.

Figures 4a, b, and c show the effect of the slight change in buyer strategies. Indeed, the fraction of buyers playing ALL C decreases and the strategy RT 25 C is more successful in this regard. However, it is not self-evident that the success of RT 25 C implies a higher rate of trust and cooperation. RT 25 C, with its single unconditional cooperation is not contingent on confiding buyers playing ALL C and, at the same time, is more successful because not exploitable. Since ALL C fails to survive in the buyer population, the deceitful seller strategy ALL D is not able to persist either. What is left, is a functioning market with trustful but cautious buyers and trustworthy sellers.

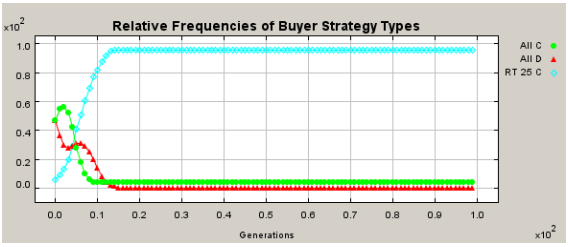

Figure 4a: Fractions of strategy types in the buyer population for 100 generations

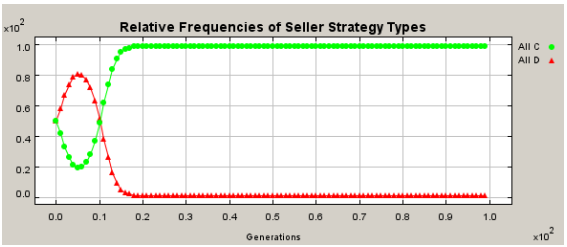

Figure 4b: Fractions of strategy types in the seller population for 100 generations

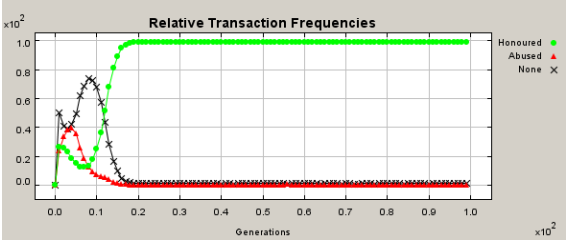

Figure 4c: Fractions of transaction types for 100 generations

The scenario experiments show that, despite the lack of an external sanctioning authority, cooperation evolves under two conditions. On the one hand, a minimal fraction of buyers must make use of the sellers' reputation in their buying strategies. On the other hand, trustworthy sellers must be given opportunities to gain a good reputation through their cooperative behavior. Strategies which account for both these requirements turn out to be successful and conducive to a market based on trust and cooperation at the same time.

Tournaments

By organizing round-robin tournaments, we aim to find new seller and buyer strategies that in competing with each other challenge our artificial market system. Strategies were collected at four different occasions and were mostly handed in by students or graduates from different universities and faculties. In every tournament, the best buyer and seller strategy could win a book token that amounted to 20 Euros. The rules to the tournament were told in writing and contained virtually the same information as the section describing the simulation experiments. Figures 5a, b, and c, show a typical run of a tournament with 11 buyer and 22 seller strategies – all strategies that have been handed in so far.

KELLER, the winning seller strategy (see Figure 5b), rewards trust three times in a row and abuses it twice subsequently. Then it increases its cooperative and fraudulent periods by one. From now on it repeats this behavior persistently. KELLER's rate of cooperation is at 60% in the beginning and decreases continuously to 50%. The winning strategy in the buyer population (see Figure 5a), SKOPEKB, is an expected value strategy. It takes the reputation index of the seller as the seller's willingness to cooperate and gives trust only if its expected payoff from a transaction with that seller is greater than or equal to the payoff for not giving trust ($P \leq rR + (1\text{-}r)S$). Solving for r reveals that SKOPEKB corresponds to a RT-strategy with a reputation threshold at 1/3. Note that SKOPEKB always gives trust to KELLER since KELLER's reputation index never falls below 0.5.[3]

These results are similar to those we obtained in many other constellations of buyer and seller strategies. In the seller population, those strategies which first build a reputation good enough to attract trustful buyers and exploit these buyers subsequently are the most successful ones. Accordingly, in the buyer population, strategies which apply the expected value principle in order to evaluate the trustworthiness of their interaction partner and to decide whether to give trust or to refuse it on the basis of that evaluation seem mostly to succeed.

In further experiments, the successive introduction of RT-strategies with an ever higher reputation threshold indeed reduced the fraction of fraudulent transactions. However, above a certain level, RT-strategies died out too early and other more forgiving buyer strategies were successful.

[3] Besides SKOPEKB, the strategy JANN applies the expected value principle. The strategy STADELMANNB which corresponds to a RT-Strategy with a reputation threshold at 1/3 is virtually identical to SKOPEKB and therefore is not considered here.

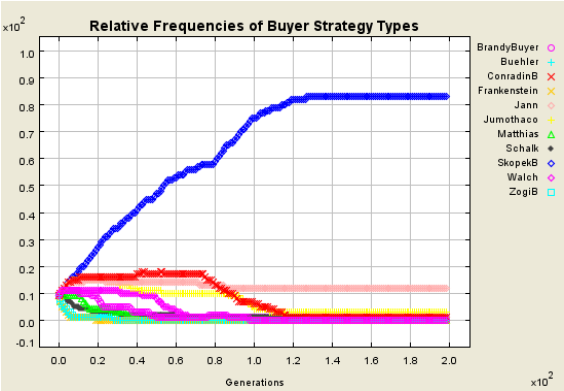

Figure 5a: Fractions of strategy types in the buyer population for 200 generations

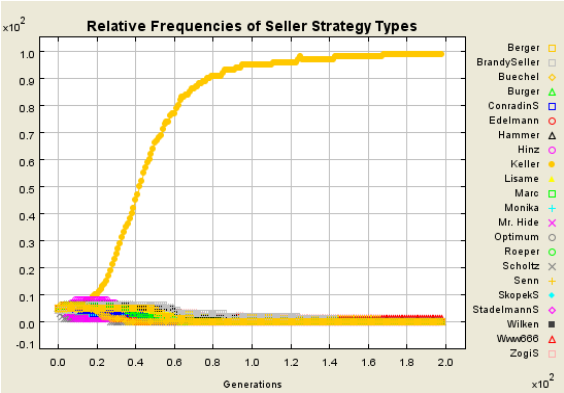

Figure 5b: Fractions of strategy types in the seller population for 200 generations

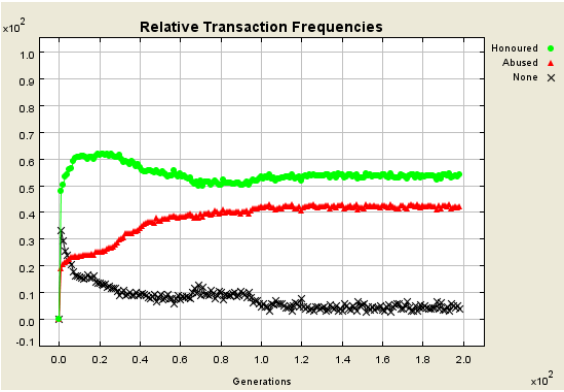

Figure 5c: Fractions of transaction types for 200 generations

Discussion

Despite the availability of data on the internet, the question of whether a reputation system alone allows an internet auction market to function can hardly be addressed by its analysis. The extent to which the operator of an internet-auction platform is able to keep the market participants from behaving opportunistically cannot be observed or manipulated. Therefore, we research this question with simulation experiments. Our agent-based model consists of a buyer and a seller population with agents applying different buyer and seller strategies. Seller and buyer agents meet several times in one generation. At every interaction they play the trust game with the seller as the second mover. If a transaction took place, the buyer rates the seller. The ratings determine a seller's reputation index which can be accounted for by buyers in subsequent interactions with that seller. However, repeated interactions between the same two agents are ignored in our setting. Through ecological analysis we explore different strategies in terms of their success and their contribution to supporting or impeding trust and cooperation.

Our scenario experiments reveal that trust and cooperation can evolve under two conditions. Some minimal fraction of buyers must make use of the sellers' reputation in their buying strategies and trustworthy sellers must be given opportunities to gain a good reputation through their cooperative behavior. Buyer strategies which account for both these requirements turn out to be successful and conducive to a market based on trust and cooperation. None the less, our artificial market is not immune against deceitful sellers. The results from round-robin tournaments with many different buyer and seller strategies show that seller strategies which first build up a good reputation and exploit trustful buyers subsequently are successful. Even buyer strategies which only cooperate with sellers with a high reputation are no remedy for deceitful sellers since other less restrictive buyer strategies are more successful. However, after all, buyer strategies which apply the expected value principle succeed in the buyer population.

Our results therefore suggest that cooperation can be established to a large extent – but not completely – in the absence of an external enforcer. A small number of deceitful sellers are able to hold their ground even in the presence of a reputation mechanism. These findings deviate from those produced by the tournaments conducted by Robert Axelrod (1984) in that they provided evidence for niches in which deceitful strategies can survive and persist.

Bibliography

Axelrod, R. 1984, *The Evolution of Cooperation*, New York: Basic Books

Berger, R. and Schmitt, K. 2005, "Vertrauen bei Internetauktionen und die Rolle von Reputation, Informationen, Treuhandangebot und Preisniveau". In *Kölner Zeitschrift für Soziologie und Sozialpsychologie 57*, pp. 86-111

Bolton, G.E., Katok, E. and Ockenfels, A. 2004, "Trust among Internet Traders: A Behavioral Economics Approach". In *Analyse & Kritik 26*, pp. 185-202

Bolton, G.E., Katok, E. and Ockenfels, A. 2005, "Cooperation among Strangers with Limited Information about Reputation". In *Journal of Public Economics 89*, pp. 1457-1468

Brinkmann, U. and Meifert, M. 2003, "Vertrauen bei Internet-Auktionen. Eine kritische Stellungnahme". In *Kölner Zeitschrift für Soziologie und Sozialpsychologie 55*, pp. 557-565

Buskens, V. and Raub, W. 2002, “Embedded Trust: Control and Learning”. In *Group Cohesion, Trust and Solidarity 19*, pp. 167-202

Coleman, J.S. 1990, *Foundations of Social Theory*, Cambridge, MA: The Belknap Press of Harvard University Press

Dasgupta, P. 2000, “Trust as a Commodity”. In Gambetta, D., ed., *Trust: Making and Breaking Cooperative Relations*, electronic edition. Department of Sociology, University of Oxford. http://www.sociology.ox.ac.uk/papers/dasgupta49-72.pdf

Diekmann, A. and Wyder, D. 2002, “Vertrauen und Reputation bei Internet-Auktionen“. In *Kölner Zeitschrift für Soziologie und Sozialpsychologie 54*, pp. 674-693

Lahno, B. 1995, “Trust and Strategic Rationality”. In *Rationality and Society 7*, pp. 442-464

Leimar, O. and Hammerstein, P. 2001, “Evolution of Cooperation through Indirect Reciprocity”. In *Proceedings of the Royal Society London B 268*, pp. 745-753

Nowak, M.A. and Sigmund, K. 1998, “Evolution of Indirect Reciprocity by Image Scoring”. In *Nature 393*, pp. 573-577

Resnick, P. and Zeckhauser, R. 2002, “Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System”. In *The Economics of Internet and E-Commerce 11*, pp. 127-157

Snijders, C. and Zijdeman, R. 2004, “Reputation and Internet Auctions: eBay and Beyond”. In *Analyse & Kritik 26*, pp. 158-184

Wedekind, C. and Milinski, M. 2000, “Cooperation through Image Scoring in Humans”. In *Science 288*, pp. 850-852

Author biographies

Andreas Diekmann is Professor of Sociology at the Swiss Federal Institute of Technology Zurich. His areas of research are rational choice and experimental game theory, research methods and statistics, environmental sociology, demography and labor market research.

Wojtek Przepiorka is doctoral candidate in sociology at the Swiss Federal Institute of Technology Zurich. His research interests include behavioral game theory and agent-based modeling and simulation with applications in the social sciences.