Danail Dochev Istvan Simonics

Radoslav Pavlov (Eds.)

Generic Issues of Knowledge Technologies

HUBUSKA Second Open Workshop
Budapest, Hungary, 14 September 2005

Proceedings

HUBUSKA
Networking Centres of
High Quality Research on
Knowledge Technologies
and Applications

MTA SZTAKI
IMI BAS, IIT BAS
NOVITECH
UNIVERSITY KLAGENFURT

Danail Dochev
István Simonics
Radoslav Pavlov (Eds.)

Generic Issues of Knowledge Technologies

HUBUSKA Second Open Workshop
Budapest, Hungary, 14 September, 2005

Proceedings

The workshop presents the results of the project
INCO-CT-2003-003401 HUBUSKA
Specific Support Action under EU FP6 programme

Institute of Information Technologies – BAS

Table of Contents

Analysis of Trust in Electronic Markets

Sviatoslav Braynov[†‡] Radoslav Pavlov[†]

[†] Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Bl. 8 Acad. G. Bonchev Str.
Sofia, Bulgaria

[‡] Department of Computer Science
University of Illinois at Springfield
Springfield, IL 62703, USA

Abstract

Trust is important for electronic markets due to the intrinsic uncertainty and risk associated with electronic transactions. The paper presents a game-theoretic and a decision-theoretic models of trust. The game-theoretic model is applied to the problem of learning trust through repeated interactions. We show that it is not beneficial for a person always to honor trust. Instead, it could be better to alternate between honoring and abusing trust so that to keep one's trustworthiness above a certain threshold.

The paper analyzes the impact of trust on electronic market efficiency. It is shown that if economic agents hold accurate trust estimates about one another, then the social welfare, the amount of trade and the agents' utility levels are maximized. We also show that market efficiency does not require complete trustworthiness.

The paper discusses various mechanisms that make agents truthfully reveal their level of trustworthiness before the beginning of every transaction. Honest reporting at the first stage of interaction informs other agents about possible risks and helps them form realistic expectations about possible outcomes.

1 Introduction

The concept of trust is essential to transactions in complex and dynamic systems with a high degree of interdependence. In such systems, the outcomes of an action depend not only on a single actor, but also on the actions of a number of external actors and chance events. Interdependencies often introduce an element of risk of failure, damage, loss, inadvertent or malicious behavior, etc. Agents can fail to perform their tasks or to meet their commitments due to lack of incentives, lack of ability, or circumstances beyond their control. For example, a buyer may have to pay for a product before it is delivered and it may not be certain whether the product will be delivered on time. Problems could arise from events outside the control of the buyer. For instance, the seller may deliver a product of inferior quality, the product may get delayed because of transportation problems, or somebody can damage the product during transportation. In general, the actors affecting the final outcome of an action may be totaly unknown, their intentions and incentives could be difficult to predict, and their skills and knowledge could be difficult to envisage.

This naturally leads to the idea of trust. Everything that is impossible or costly to secure needs to be trusted. The concept of trust has been a subject of continuous interest in different research areas, including computer security [1], multiagent systems [2, 3, 4, 5], sociology [6], risk management [7, 8], economics and game-theory [9, 10, 11, 12].

Trust is usually considered a belief that an entity will perform in a favorable way. In other words, an actor takes the risk to depend on an entity which has partial or full control over a situation. Whether the outcome of the situation is favorable to the actor depends on the entity. For example, when a buyer is buying a second hand car, the buyer might need to trust the car dealer to sell him a good car. It is usually the case that the dealer has better information than the buyer about the quality and the condition of the car. If the buyer believes that the dealer is likely to withhold important information about the car, the buyer would certainly choose another car dealer.

Reputation is another concept [10] that is close to the notion of trust. It is worth noting, however, that the concept of reputation is more general than trust. An agent may have the reputation of being trustworthy, honest, aggressive, tough, etc. Therefore, the notion of reputation involves establishing and maintaining some individual characteristics which are publicly observable. Trust, on the other hand, could be based on private information. For example, an agent may be trustworthy without having any public reputation. Despite the difference between them, both trust and reputation can be used as a capital asset. Economic research [13] reveals that in their economic activity firms tend to convert their financial capital into reputational capital and vice versa. That is, in order to increase their reputation firms may be willing to invest some financial capital, or in order to increase their financial capital firms may be willing to sacrifice their reputation.

Reputation systems [14] have long been used as a means for measuring trustworthiness. Their use, however, is limited by several factors:

- In many cases, obtaining reputation rankings may be impossible. Internet provides vast opportunities to interact with total strangers for whom there is no history of previous transactions.
- Information obtained from reputation databases or recommender systems is usually too general to be applied to the context of a particular transaction.
- The coverage of current reputation systems is limited.
- There are problems with the aggregation of correlated reputation [15].

- Possibility for fake transactions [16].
- False identity and pseudonyms [17].
- Obtaining negative feedback from unsatisfied customers could be difficult [14]. Trust, on the other hand, could be based on private information.

2 Trust in electronic markets

The concept of trust is important to e-commerce because it affects the very essence of on-line business: the possibility to engage in a risky transaction. Internet users still fear the possibility of fraud, misuse of private information, identity change, deception, etc.

Need of trust in business transactions is usually explained by time asymmetry, lack of power, or inability to conclude perfect contracts. The *time asymmetry* argument draws on the fact that transactions are usually performed over a period of time and the actions of some agents temporally precede those of others. In other words, there is a time lag between the placement of trust and the corresponding action of the trusted party who can honor or abuse trust. For example, the delivery of goods and services by one party might occur only before the other party has made the payment. In the absence of appropriate control mechanisms, the party that acts first has to trust the other agents for fulfilling their part of the transaction.

The *contract argument* addresses the possibility of concluding definite and complete contracts between cooperating agents. The effect of legal contracts is at least threefold:

- It decreases the incentive for the trustee to abuse trust by providing appropriate sanctions.
- It compensates the trustor if trust is abused.
- It increases the trustor's expectations that the trustee will behave favorably.

Legal and economic experience [18], however, indicate that contracts usually are incomplete. Several factors contribute to contract incompleteness. The following are the most common. First, a contract may be ambiguous because the words explaining an issue are ambiguous. Second, the contract parties may fail to reach agreement about some issues, yet prefer to make a contract on the issue on which they agree. Third, the contract may have been left incomplete because the cost for the contract parties of drafting an issue may exceed the expected benefit. Finally, there may be asymmetric information about an issue because events in the world might not be mutually observable.

The *power argument* for placing trust is as follows. When an agent does not have the power to control actions of other agents (including nature) or when exerting the power is too costly for him and when other agents' actions have a bearing on his behavior or welfare, then he may be willing to place trust on the agents he cannot control. In game theory the notion of power is formalized by Harsanyi [19] The impact of power on multiagent planning is studied by Brainov and Sandholm [20]. The relation between power and trust is discussed by Luhmann [21].

It should be pointed out, however, that there is a significant difference between trust in physical markets and trust in e-commerce. Building and maintaining trust in electronic markets is more difficult without face-to-face interaction [22], partner identity, and clearly defined legal framework. In online markets it may not be possible to track down or even to identify a party in a transaction. A software agent, for example, may act on behalf of different human users at different moments at time, thereby making it difficult to relate its behavior to one physical entity.

Electronic markets also significantly reduce the costs of establishing new business contacts and changing business partners. This shortens the average life of business relationships, making it more difficult to build trust and consume the benefits of a long trust-based relationship.

Another factor that adversely affects trust is the dynamics and volatility of electronic markets. In e-commerce trust can be destroyed in an instant by misfortune or a mistake and the effect of it could be globalized throughout many interconnected markets.

3 Decision-theoretic model of trust

Trust has different connotations and has been used in different meanings in different contexts by different authors. Many authors [1] consider trust to be a belief or cognitive stance that could eventually be quantified by a subjective probability. We give a brief conceptualization of trust that will help avoid confusion and will facilitate further exposition.

In modeling trust, we take a rational-choice approach. We assume that: (i) There is a preference (utility) function which measures the desirability of an outcome; (ii) Different people may have different preference (utility) functions; (iii) Trust is only possible if, for the trustor; the expected outcome of placing trust is preferred over the expected outcome of not placing trust.

We assume that trust is a bilateral relation that involves an entity manifesting trust called the *trustor* and an entity being trusted called the *trustee*. Further, we assume that

- There is an event Γ that the trustor cannot control and that depends on the trustee. The trustee may have partial or full control over Γ.
- The trustor voluntarily decides to put himself in a position dependent on Γ in the sense that the trustor will benefit if Γ occurs, otherwise he will lose.
- The trustee can honor trust by bringing about Γ, and abuse trust by not bringing about Γ. Usually, the trustee is better off by abusing trust, since bringing about Γ may incur costs and inconvenience to the trustee.

In a trust relation, the trustor takes the risk to depend on the trustee for a certain event the outcome of which depends completely or partially on the trustee. We assume that trustworthiness could be measured by the probability of Γ. For example, the trustee could be a user and $\Gamma = \{$The user does not attempt to elevate his privileges$\}$. Another interpretation is $\Gamma = \{$The quality of the software delivered by the trustee meets the trustor's expectations$\}$.

Consider a trustor who uses a software product developed by the trustee. If the trustor cannot control the process of the design, implementation, and installation of the product, then the trustor may take the risk to trust the developer that the product meets its specifications. It is usually the case that the product developer has better information than the trustor about the security and reliability of the product. A problem may arise if the trustee decides to withhold information from the trustor in order to save production costs and deliver a product of inferior quality. Here, the trustee deliberately and intentionally decides to abuse trust. Moreover, additional problems may occur from events outside the control of the trustee and the trustor. For example, the product may contain a third party off-the-shelf component which happens to be unstable. Here, the problem does not arise because of untrustworthy behavior on the part of the trustee. In another scenario, the trustee is completely willing to deliver a high quality product and puts a lot of effort in the program implementation and testing. Unfortunately, because of lack of experience the trustee overlooks established security practices and delivers an insecure product.

In general, trust can be divided into two major categories:

- Trust based on good will: the trustor believes that the trustee has the good will not to abuse trust. That is, the trustee will bring about the favorable for the trustor event Γ even though it may not be beneficial for the trustee. In this case, the trustee has always the power and capacity (the full control over Γ) to honor trust. Whether he will honer trust depends on his incentives and benefits.

- Trust based on competence: the trustor believes that the trustee is competent enough to bring about the favorable event Γ. It could be the case that the trustee is willing, but not competent enough to honor trust (to bring about Γ).

Trustworthiness can also be classified as *perceived* or *actual*. Perceived trustworthiness is defined as the trustor's subjective belief in the competence or the good will of the trustee. Obviously, the perceived trustworthiness could be different from the objective trustworthiness, which is the actual incentive or capacity of the trustee. Perceived trustworthiness is measured with the thrustor's subjective probability of Γ, and the actual trustworthiness is measured with the objective probability of Γ. For example, a user may install a program, believing that the program failure rate is $\hat{\theta}$, while the actual failure rate is θ.

It should be noted that the concept of trust has two main meanings which are often confused. Trust could refer to both the trustfulness of the trustor and the trustworthiness of the trustee, where the *trustfulness* of the trustor is the extent to which the trustor places trust in the trustee, whereas the *trustworthiness* of the trustee is the extent to which the trustee honors trust. It is obvious that the trustfulness of the trustor largely depends on his expectation about the trustee's trustworthiness (the perceived or subjective trustworthiness). The higher the perceived trustworthiness, the more wiling the trustor is to place trust in the trustee.

In order to form a trusting belief, the trustor must have enough knowledge about the trustee's incentives or competence. Such knowledge usually comes from two sources: individual and public. Individual sources include the history of previous interactions with the trustee, recommendations of other agents, etc. Public sources usually include the publicly established reputation of the trustee. In some cases, public and individual knowledge could be conflicting. For example, based on a series of positive experiences with a product developer, the trustor may decide to trust him even though it is well known that the developer has incentives to deliver inferior products.

The concept of trust is different from concept of reputation which relates to publicly established and recognized features of an entity. For example, an agent could be completely trustworthy without having any public reputation. Such an agent could efficiently interact with a few other agents that trust him and know him for a long time. Establishing public reputation, however, could be costly or could require long time.

In our interpretation of trust, the trustor does not have control over the trustee's behavior and the event Γ. In a sense, trust and control are mutually exclusive. The more control the trustor has over the trustee, the less the trustor needs to place trust. Obviously, trust relates to the state of being dependent, while control relates to the ability to keep someone in a dependent position.

Formally, the trustor's utility function can be denoted by:

$$U(\hat{\theta}, \Gamma(p_1, ..., p_n)) \quad (1)$$

where U is the trustor's utility, $p_1, ..., p_n$ are parameters describing the event Γ, and $\hat{\theta}$ is the degree of perceived trustworthiness, i.e., the probability with which Γ is expected to happen. A natural measure for the perceived trustworthiness is the subjective probability of Γ, i.e., the trustor's strength of belief that the trustee will bring about Γ. The degree of perceived trustworthiness $\hat{\theta}$ could differ from the degree of actual trustworthiness θ.

The event Γ is favorable to the trustor:

$$\frac{\partial U(\theta, \Gamma(p_1, ..., p_n))}{\partial \theta} \geq 0$$

That is, the trustor benefits from higher trustworthiness. The case of complete trustworthiness is represented by $\theta = 1$, and vice versa, the trustee is completely untrustworthy when $\hat{\theta} = 0$:

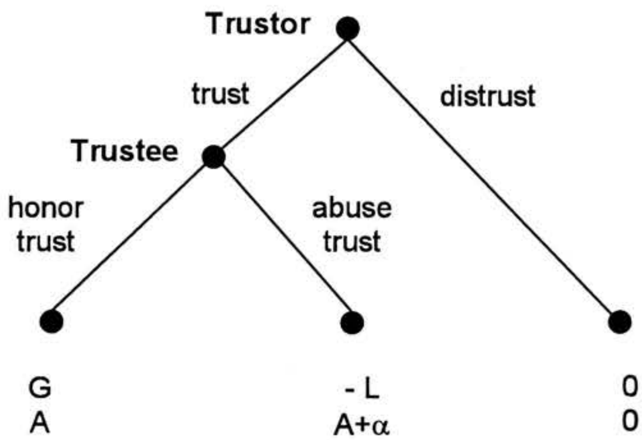
$$U(1, \Gamma(p_1, ..., p_n)) > 0$$
$$U(0, \Gamma(p_1, ..., p_n)) < 0$$

If we assume that utility is a continuous function of trustworthiness, then there is a threshold level θ_0, $\theta_0 \in [0, 1]$, that separates trustworthiness from untrustworthiness:

$$U(\theta, \Gamma(p_1, ..., p_n)) \geq 0 \qquad \text{for all} \qquad \theta \geq \theta_0$$

The trustor is always better off if the other agent's trustworthiness exceeds the threshold θ_0 which depends on the event Γ and its parameters $p_1, ..., p_n$.This defines a natural *participation constraint*: the trustor will place trust on the trustee (or will voluntarily agree to depend on the trustee) if the trustee's perceived trustworthiness exceeds θ_0. The participation constraint corresponds to the intuition that an agent will only engage in an interaction if the trustworthiness of the other party exceeds some threshold (the level of acceptable trustworthiness), which depends on the interaction context (through parameters $p_1, ..., p_n$) and on the trustor (through the trustor's utility function U). In

Figure 1: Extensive form of a Trust Game



other words, the threshold θ_0 is both objectively and subjectively determined. Such a formalization agrees with the threshold model of trust (Kee and Knox,1970; Coleman, 1990; Tan and Thoen, 1999).

Such a formalization of trust is domain independent and captures a wide range of applications, where the trustor believes that the trustee will behave in some expected way specified by the event Γ. Depending on the context the event Γ, can be given different interpretations.

4 Game-theoretic model of trust

A trust relation can be modeled as a one-shot extensive-form game shown in Figure 1. The game starts with a move made by the trustor, who must choose between placing trust and not placing trust. If the trustor decides not to place trust, the game is over and both the trustor and the trustee receive a payoff 0. If the trustor places trust, the trustee has two choices: to honor or to abuse trust. The trustee is usually better off by abusing trust (by not bringing about Γ), while the trustor is better off if trust is honored, i.e., Γ is brought about. If the trustee honors trust, the trustor and trustee receive G and A, respectively. If the trustee abuses trust, the trustor receives $-L$ and the trustee receives $A + \alpha$. Here, α represents the incentive of the trustee to abuse trust. We assume that these payoffs represent utilities for the trustee and the trustor that correspond to the outcomes of the game. The relationship between payoffs is as follows: $G > 0, L > 0, A > 0, \alpha > 0$.

If the trustor places trust, the trustee chooses between A and $A + \alpha$, so surely he will abuse trust. Anticipating this, the trustor chooses between lost L if he places trust and 0 if he does not. Therefore, there is a Nash (and subgame-perfect) equilibrium of the game in which the trustor does not place trust, and the trustee always abuses trust. The equilibrium is socially inefficient because both the trustor and the trustee are worse off in the situation in which trust is not placed than in the case where trust is placed and honored.

This is one equilibrium of the game, but there is another in which the trustor places trust if the trustee honors trust, and the trustee always honers trust. The problem with this equilibrium is that it is not as credible as the first one. It would be irrational for the trustee to honor trust once the trustor has chosen to place trust. In other words, the second equilibrium is not subgame-perfect.

Which equilibrium will be chosen depends on the trustor's expectation of the trustworthiness of the trustee (the so called perceived trustworthiness). If the game is played only once and does not affect the interactions that the trustee is going to have with other agents, then the more plausible equilibrium is the non-cooperative one in which trust is withhold.

5 A Model of Trust Learning

In this section we use the game-theoretic model of trust developed in the previous section to analyze how trust could be learnt over time. A detailed description of the model can be found in [23].

Research on risk perception [7] indicates that trust typically is learned gradually over many repeated interactions. As we have seen in the previous section, one interaction is not sufficient to establish trust because the trustee has a short-run incentive to abuse trust. Hence, the non-cooperative equilibrium prevails.

In this section, we consider repeated interactions between the trustee and the trustor. A repeated game consists of infinite repetitions of the Trust game, shown in Figure 1. A repeated game allows the trustor to modify his expectations about the incentives and abilities of the trustee by observing his behavior in past interactions. Initially, the trustor starts with some prior beliefs about the trustworthiness of the trustee. As the game proceeds, the beliefs are adjusted. After a series of positive interactions, the trustor's trust in the trustee will gradually increase, and vice versa, multiple unsuccessful interactions in which trust is abused, may lead to such a low estimate of the trustworthiness of the trustee that the trustor no longer wishes to interact. In this case the game ends. Let θ_0 be the threshold, such that whenever the perceived trustworthiness of the trustee, $\hat{\theta}$, falls bellow θ_0, the trustor exits the game. This defines a natural *participation constraint*: the trustor will interact with the trustee (or will voluntarily agree to depend on the trustee) if the trustee's perceived trustworthiness exceeds the threshold:

$$\hat{\theta} \geq \theta_0$$

The structure of the game and the payoffs are common knowledge between the trustor and the trustee. To introduce reputation effects into the game, suppose that the objective trustworthiness of the trustee is θ. In other words, the trustee abuses trust with probability θ, if trust is placed. Further, assume that θ is private knowledge known only to the trustee. Therefore, the trustor will place trust if the trustor's expected benefit from placing trust exceeds the benefit from withholding trust:

$$\hat{\theta}G - (1 - \hat{\theta})L > 0$$

where $\hat{\theta}$ is the perceived trustworthiness of the trustee, i.e, the trustor's estimate of θ. Therefore, the threshold is:

$$\theta_0 = \frac{L}{L+G}$$

In this setting, the problem of learning trust is reduced to the problem of estimating θ based on a series of interactions in which trust is abused or honored. In order to preserve belief consistency, it is natural to assume that the trustor performs belief revision using Bayes' rule. To find a Bayesian estimator of the trustworthiness of the trustee, we use the Laplace succession rule [24]. The rule estimates the probability that the trustee will honor trust in the next interaction, given that he honored trust in h out of k interactions, a more general estimator can be used:

$$\hat{\theta} = \frac{h+1}{k+2} \tag{2}$$

Formula 2 suggests that there is a Nash equilibrium of the repeated Trust game in which:

- The trustor places trust if and only if:

$$\hat{\theta} = \frac{h+1}{k+2} > \theta_0$$

- The trustee alternates between honoring and abusing trust so that to keep the trustor's estimate $\hat{\theta}$ above the threshold θ_0.

The next Proposition follows naturally from the definition of the repeated Trust game and from players' equilibrium strategies:

Proposition 1 *In the equilibrium, the trustee alternates between honoring and abusing trust so that:*

$$\frac{h}{a} \geq \frac{L}{G}$$

where a is the number of times trust is abused, h is the number of times trust is honored, L is the lost to the trustor if trust is abused, and G is the gain to the trustor if trust is honored.

Proposition 1 suggests that the ratio of the number of times trust is honored to the number of times trust is abused depends on the ratio of the loss to the gain for the trustor. The result is intuitively reasonable. The more the trustor can lose in the game, the more frequently the trustee honors trust. Since by definition $L > G$, the trustee needs to honor trust more frequently than to abuse trust.

6 The effect of trust on market efficiency

In this section, we study the effect of trust on market efficiency. More specifically, we show that if agents hold accurate trust estimates about one another, then the social welfare, the amount of trade and the agents' utility levels are maximized. We also show that market efficiency does not require complete trustworthiness. Untrustworthy agents could transact as efficiently as trustworthy agents, provided that they hold accurate estimates of one another. Therefore, what really matters is not the actual level of trustworthiness, but the accuracy of individual estimates. A market in which agents are trusted to the degree they deserve to be trusted is as efficient as a market with complete trustworthiness.

A detailed description of our analysis can be found in [25, 3]. The analysis is done within the framework of a bilateral negotiation involving a buyer and a seller. The seller produces some commodity and sells it to the buyer. We assume that the seller always delivers the commodity i.e., he is completely trustworthy. The buyer's trustworthiness, however, may vary. In other words, the buyer pays with some probability θ, $\theta \in [0,1]$. We assume that the seller delivers first and after that the buyer pays. In this case, the seller depends on the buyer for the event $\Gamma = \{\text{the buyer pays}\}$, and the seller has to choose whether to enter into a transaction without being able to control Γ.

Let $\hat{\theta}$ and θ denote respectively the buyer's perceived and actual trustworthiness. That is, $\hat{\theta}$ is the seller's estimate of θ. One way to look at θ is to see it as an indicator of the buyer's willingness to pay. In another interpretation, θ could be the buyer's ability or capacity to pay (the buyer, for example, may be willing to pay, but may not have available funds). θ could also be the probability that the contract between the buyer and the seller can be enforced by an enforcement agency. That is, θ could be the probability of detecting particular violation, imposing particular type of sanctions, and using particular type of procedures for adjudicating disputes.

The seller's estimate, $\hat{\theta}$, could differ from the buyer's actual trustworthiness, θ. It is usually the case that the buyer is undertrusted. That is, $\hat{\theta} < \theta$. For example, many risk assessment firms treat the lack of credit history as a lack of trust. This is usually motivated by the fact that the marginal cost of obtaining additional evidence of trust exceeds the marginal benefit of the evidence. Undertrusting is a typical example of a market inefficiency produced by inaccurate trust estimates. In the extreme case, undertrusting could produce such a low estimate of the partner's trustworthiness, that an agent might decide not to participate in a transaction, even when the partner is completely trustworthy.

The following proposition shows the optimal amount of trust necessary to achieve market efficiency.

Proposition 2 *When the trust matches trustworthiness, $\hat{\theta} = \theta$, the seller and the buyer both maximize their individual utility functions. Moreover, the social welfare is maximized and the maximal possible output is produced and sold.*[1]

The next proposition shows that the undertrusting leads to market inefficiencies.

Proposition 3 *When the buyer is undertrusted, $\hat{\theta} < \theta$, the social outcome is not optimal. Namely, quantity exchanged and the utilities of both agents are smaller than those obtained in the case when the seller places the correct amount of trust on the buyer, $\hat{\theta} = \theta$.*

[1] The proofs of Propositions 2 and 3 can be found in [25]

According to Propositions 2 and 3, the case where the seller trusts the buyer to the extent that the buyer deserves to be trusted is optimal for society. By an optimal outcome we mean one that maximizes the social welfare, the quantity produced, and the agents' utility functions. In order to obtain efficiency it is not necessary that the buyer be trustworthy. The only relevant factor is the seller's accuracy in estimating the buyer's trustworthiness. Any underestimation of the buyer's trustworthiness tends to harm each agent and society as a whole.

It is worth pointing out that in the case when trust matches trustworthiness, $\hat{\theta} = \theta$, individual interests coincide with the social interest in the sense that agents maximize their utilities, maximizing at the same time the social welfare.

If the buyer is distrusted, $\hat{\theta} < \theta$, he may try to convince the seller of his trustworthiness. One possible way for a distrusted buyer to signal his trustworthiness is to make an advance payment to the seller, i.e., to pay some amount before the seller delivers the commodity. We assume that in the absence of payment the seller is relieved from his obligations. That is, if the buyer does not make an advance payment after he has promised to do so, the seller is relieved from his obligation to deliver.

Proposition 4 *If the buyer is distrusted, $\hat{\theta} < \theta$, and the agents choose an advance payment contract, then the social welfare is maximized. The quantity exchanged with an advance payment contract equals the quantity exchanged with an uncertain payment contract. Moreover, the advance payment contract gives each agent higher utility than the uncertain payment contract.*

According to Proposition 4 if the buyer is trustworthy, he should pay the entire price in advance. This corresponds to our intuition that paying before or after the delivery does not make a difference for a trustworthy buyer. Furthermore, Proposition 4 says that the advance payment contract is better than an uncertain payment contract in that it provides agents with higher utility levels.

From Propositions 4 it follows that when trustworthiness is underestimated, the advance payment contract is optimal. Advance payment contracts could also serve as a screening device. That is, they could help separate trustworthy agents from untrustworthy ones. If the buyer is trustworthy, he should not object to paying the whole price in advance. Therefore, a buyer who declines an advance payment contract is not as trustworthy as he claims to be.

7 Incentive Compatible Mechanism for Trust Revelation

In this section, we discuss a solution to the problem of obtaining trust estimates: an incentive-compatible mechanism in which agents truthfully reveal their trustworthiness at the beginning of every interaction. In such a mechanism agents always report their true level of trustworthiness, even if they are untrustworthy. Honest reporting at the first stage of interaction lets other agents know the interaction risk and form realistic expectations about possible outcomes. The mechanism works for both single-step and repeated interactions. A complete description of the mechanism is given in [4, 26]

To illustrate an incentive-compatible mechanism for trust revelation consider the following scenario. The trustor and the trustee take part in a transaction, $T(q,\theta)$, which can be described by a generic parameter q and the level of the trustworthiness of the trustee θ. The generic parameter q denotes the transaction terms which depending on the transaction context, could include price, quantity, services provided, or combination of them. The parameter θ denotes the trustworthiness of the trustee, i.e., the probability that the trustee will carry out his part of the transaction.

We assume that the trustor always executes the transaction i.e., he is completely trustworthy. The trustee's trustworthiness, however, may vary. In other words, the trustee carries out the transaction with some probability θ, $\theta \in [0,1]$. The level of trustworthiness θ is a private value known only to the trustee. In other words, in the beginning of the transaction the random realization of the trustworthiness of the trustee is not observable by the trustor. Further, we assume that the trustor offers a transaction to the trustee by specifying the terms of the transaction q. The trustee could accept or reject the offer of the trustor.

Given this setting, what is the optimal transaction for the trustor to offer? In other words, the trustor faces the problem of designing an optimal (i.e, utility maximizing) offer. For example, the trustor might ask the trustee to declare his level of trustworthiness θ, and then based on the declaration, the trustor could choose the value of q. It is usually the case, however, that the trustee has a strong incentive to misrepresent his level of trustworthiness. By declaring more trustworthiness the trustee usually enjoys more benefits of future cooperation and more opportunities to abuse trust. By implementing an appropriate incentive-compatible mechanism the trustor can make the trustee truthfully reveal his level of trustworthiness.

An incentive-compatible mechanism for trust revelation works in the following way. First, at the beginning of the interaction the trustee declares his trustworthiness, $\tilde{\theta}$, and after that the transaction terms q are chosen by the trustor as a function q, $q = q(\tilde{\theta})$, of the trustee's declaration. In an incentive-compatible mechanism the trustee finds it optimal to report his trustworthiness truthfully, i.e., $\theta = \tilde{\theta}$.

The following proposition establishes the existence of an incentive-compatible mechanism for trust revelation.

Proposition 5 *There exists a single-transaction incentive-compatible mechanism $q = q(\tilde{\theta})$ that makes the trustee truthfully reveal his trustworthiness at the beginning of the transaction, if the following conditions hold:*

$$\frac{\partial^2 U(q,\tilde{\theta})}{\partial q^2} < 0$$

$$\frac{\partial U(q^*,\tilde{\theta})}{\partial q} = 0$$

for some q^*, and

$$\frac{\partial^2 U(q^*,\tilde{\theta},)}{\partial q \partial \tilde{\theta}} \neq 0$$

where $U(q,\theta)$ is the trustee's utility function.[2]

[2]The proofs of Propositions 2 and 3 can be found in [4, 26]

The basic idea behind the mechanism is that the trustee achieves maximum utility only if specific transaction terms are chosen. Since the transaction terms, q, depend on the trustee's declaration, $\tilde{\theta}$, by declaring his true level of trustworthiness the trustee maximizes his utility. The mechanism is designed in a way to give the trustee sufficient incentives to report truthfully even if it is untrustworthy.

The results of Proposition 5 can be carried over to the case of repeated transactions. By a repeated transaction between the trustor and the trustee we mean finitely or infinitely many repetitions of the single transaction in which both the trustor and the trustee discount their future payoffs at a constant rate.

In the case of a repeated transaction, it is not obvious that the trustee will truthfully declare his trustworthiness. For example, the trustee could lie in the first transaction in order to obtain more favorable transaction terms, q, in the subsequent transactions. the next proposition shows that trust revelation is possible in the case of a repeated transaction.

Proposition 6 *Under the conditions of Proposition 5, for any repeated transaction between the trustor and the trustee there exists an incentive-compatible interaction mechanism* $q = q(\tilde{\theta})$ *that makes the trustee truthfully reveal his trustworthiness in the beginning of the first transaction.*

An incentive-compatible mechanism for trust revelation provides several advantages. It does not rely on a third party for providing information of for backing up a transaction. It does not depend on collecting and analyzing information about untrustworthy agents. In addition, it does not require the estimation of other agents' trustworthiness. This solves the problem of inaccuracy of individual estimates and avoids many inefficiencies caused by inconsistent or inaccurate estimates. The mechanism also eliminates the need to speculate about other agents' intentions and beliefs.

8 Conclusions

The paper defines a game-theoretic and a decision-theoretic model of trust. The game-theoretic model is used to show how trust can be learnt over time. The equilibrium analysis demonstrates that instead of always honoring trust, the trustee alternates between honoring and abusing trust so that to keep the trustor's estimate above a certain threshold.

The decision-theoretic model of trust is used to analyze impact of trust on electronic markets. The paper shows how trust affects market efficiency, social welfare, the volume of trade, and the profits of market participants. The paper also demonstrates that complete trustworthiness is not necessary for market efficiency. Instead, every transaction has its optimal level of trustworthiness that is needed for the transaction completion.

The paper introduces the concept of trust-based mechanism design. The primary objective of trust-based mechanism design is to develop and implement optimal markets in which every agent discloses truthfully his/her trustworthiness before a transaction starts.

Most existing trust-building mechanisms assume complete or partial intervention of trusted third parties, such as enforcing agencies, reputation systems, etc. Such arrangements are usually costly and not always available to economic agents. The primary advantage of the trust-based mechanism design is that the mechanism is self-enforcing and does not rely on third parties, i.e., it is in the best interest of every agent to declare his/her true level of trustworthiness.

References

[1] A. Josang, "Trust-based decision making for electronic transactions," in *Proceedings of the Fourth Nordic Workshop on Secure Computer Systems (NORDSEC '99)*, Sweden, 1999, pp. 496–502.

[2] S. Marsh, *Formalizing Trust As a Computational Concept*, Ph.D. thesis, University of Stirling, U.K., 1994.

[3] S. Brainov and T. Sandholm, "Contracting with uncertain level of trust," in *Proceedings of the First ACM Conference on E-Commerce*, Denver, Colorado, 1999, pp. 15–21.

[4] S. Brainov, "An incentive compatible trading mechanism for trust revelation," in *Proceedings of the IJCAI'01 Workshop on Economic Agents, Models and Mechanisms*, 2001, pp. 62–70.

[5] C. Castelfranchi and R. Falcone, "The dynamics of trust: From beliefs to action," in *Proceedings of the Second Workshop on Deception, Fraud and Trust in Agent Societies*, Seattle, 1999, pp. 41–54.

[6] J. Coleman, *Foundations of Social Theory*, Harvard University Press, 1990.

[7] P. Slovic, "Risk perception and trust," in *Trust: Fundamentals of Risk Analysis and Risk Management*, Vlasta Molak, Ed. Lewis Publishers, 1997.

[8] S. Kaplan B. Garrik, "On the quantitative definition of risk," *Risk Analysis*, vol. 28, pp. 11–27, 1981.

[9] P. Dasgupta, "Trust as a commodity," in *Making and Breaking Cooperative Relations*, D. Gambeta, Ed. Basil Blackwell, 1990.

[10] D. Kreps and R. Wilson, "Reputation and imperfect information," *Journal of Economic Theory*, vol. 27, pp. 253–279, 1982.

[11] C. Snijders, *Trust and Commitments*, ICS, 1996.

[12] V. Buskens, *Social Networks and Trust*, Kluwer, 2002.

[13] A. Boot, S. Greenbaum, and V. Thakor, "Reputation and discretion in financial contracting," *American Economic Review*, vol. 83, pp. 1165–1183, 1993.

[14] P. Resnick, R. Zeckhauser, F. Friedman, and K. Kuwabara, "Reputation systems," *Communications of the ACM*, vol. 43, no. 12, pp. 45–48, 2000.

[15] M. Schillo, P. Funk, and M. Rovastos, "Who can you trust: Dealing with deception," in *Proceedings of the Second Workshop on Deception, Fraud and Trust in Agent Societies*, Seattle, 1999, pp. 95–106.

[16] Zacharia and Maes, "Trust management through reputation mechanisms," *Applied Artificial Intelligence*, vol. 14, no. 8, pp. 881–907, 2000.

[17] E. Friedman and P. Resnick, "The social cost of cheap pseudonyms," Working paper, 1998.

[18] L. Werin and H. Wijkander, *Contract Economics*, Basil Blackwell, 1992.

[19] J. Harsanyi, "Measurement of social power, opportunity costs and the theory of two-person bargaining games," *Behavioral Science*, vol. 7, pp. 67–80, 1962.

[20] S. Brainov and T. Sandholm, "Power, dependence and stability in multiagent plans orlando, 1999.," in *Proceedings of the National Conference on Artificial Intelligence AAAI'99*, Orlando, Florida, 1999, pp. 11–16.

[21] N. Luhmann, *Trust and Power*, John Wiley and Sons, 1979.

[22] E. Rocco, "Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact," in *Proceedings of ACM CHI 98 Conference on Human Factors in Computing Systems*, 1998, pp. 496–502.

[23] S. Brainov, "Trust learning based on past experience," in *Proceedings of the IEEE International Conference on Integration of Knowledge Intensive Multiagent Systems (KIMAS)*, 2005, pp. 197–201.

[24] W. Feller, *In Introduction to Probability Theory and Its Applications*, vol. 1, John Wiley and Sons.

[25] S. Braynov and T. Sandholm, "Contracting with uncertain level of trust," *Computational Intelligence*, vol. 14(4), pp. 501–514, 2002.

[26] S. Braynov and T. Sandholm, "Trust revelation in multiagent interaction," in *in Proceedings of CHI'02 Workshop on The Philosophy and Design of Socially Adept Technologies*, 2002, pp. 55–60.