"CONFESS". Eliciting Honest Feedback without Independent Verification Authorities

Radu Jurca and Boi Faltings

Artificial Intelligence Laboratory (LIA),
Swiss Federal Institute of Technology (EPFL),
CH-1015 Lausanne, Switzerland
{radu.jurca, boi.faltings}@epfl.ch
http://liawww.epfl.ch/

Abstract. Reputation mechanisms offer an efficient way of building the necessary level of trust in electronic markets. In the absence of independent verification authorities that can reveal the true outcome of a transaction, market designers have to ensure that it is in the best interest of the trading agents to report the behavior in transactions truthfully. As opposed to side-payment schemes that correlate a present report with future reports submitted about the same agent, we present a mechanism we have called "CONFESS", that discovers (in equilibrium) the true outcome of a transaction by analyzing the two reports coming from the agents involved in the exchange. For two long-run rational agents, we show that it is possible to design such a mechanism that makes cooperation a stable equilibrium.

1 Introduction

The availability of ubiquitous communication through the Internet is driving the migration of business transactions from direct contact between people to electronically mediated interactions. People interact electronically either through human-computer interfaces or even through programs representing humans, so-called agents. In either case, no physical interactions among entities occur and the systems are much more susceptible to fraud and deception.

Traditional methods to avoid cheating, involving strong cryptography and Trusted Third Parties (TTP's) that overlook every transaction, are very costly and sometimes even impossible to apply due to the complexity and heterogeneity of the environment. Moreover, network communities often have a strong desire of being independent of any authorities, as illustrated by the successful P2P systems.

Reputation mechanisms offer a novel and efficient way of ensuring the necessary level of trust in electronic markets. They are based on the observation that agent strategies change when we consider that interactions are repeated: the other party will remember past cheating, and changes its terms of business accordingly in the future. In this case, the expected future gains due to future

transactions in which the agent has a higher reputation can offset the loss incurred by not cheating in the present transaction. This effect can be amplified considerably if such reputation information is shared among a large population and thus multiplies the expected future gains made accessible by honest behavior.

Theoretic research on reputation mechanisms started with the seminal papers of Kreps, Milgrom, Wilson and Roberts [10–12] who explained how a small amount of incomplete information is enough to generate the *reputation effect*, (i.e. the preference of agents to develop a reputation for a certain *type*) in the finitely repeated Prisoners' Dilemma game and Selten's Chain-Store game [15].

Fudenberg and Levine [6] and Schmidt [14] continue on the same idea by deriving lower bounds on the equilibrium payoff received by the reputable agent in two classes of games in which the reputation effect can occur.

A number of computational trust mechanisms have been developed based on both direct (i.e. interaction-derived) and indirect (i.e. reported by peers) reputation information [1, 2, 16]. This class of mechanisms, however intuitive, are ad-hoc, do not provide rational participation incentives, and impose restrictions on the acceptable behavior of the agents.

In [4] Dellarocas presents an efficient binary reputation mechanism that encourages a cooperative equilibrium in an environment of purely opportunistic buyers and sellers. The mechanism is centralized, it works for single-value transactions, and is robust (within certain limits) against mistakes made by reporters.

One major challenge associated with designing reputation mechanisms is to ensure that truthful reports are gathered about the actual outcome of the transaction. In a typical trading interaction, e.g. an exchange between a seller (he) and a buyer (she), the buyer is required to first pay and then wait for the purchased good to be shipped to the intended destination. While the payment of the buyer can be easily verified with the authority intermediating the transaction (e.g. the credit card company), it is very difficult to verify that the seller has indeed shipped the promised good. We start from a typical assumption about online environments: the outcome of one transaction (i.e. the seller has shipped or not the good) is only known to the parties involved. Any reputation mechanism will therefore have information that is distorted by the strategic interests of the reporters.

Most real situations do not make it rational for an agent to report the truth. The private information of a buyer for example, about the trustworthiness of a seller is often regarded as an asset which should not be freely shared. Paying for the buyer's reputation report could overcome this inconvenient, however, no guarantee can be offered that the information provided is true. For example, a true positive report might create inconveniences for the reporting buyer because of decreased future availability of that particular seller. Moreover, in a competitive environment, a false negative report about a seller slightly increases the buyer's own reputation with regards to the other agents.

The problem of incentive compatibility can be addressed by paying for a reputation report, such that the payment is conditioned on the correlation with

future reports (assumed to be true) about the same seller. [13] and [8] describe such schemes that make truth revelation a Nash equilibrium. A problem with these schemes however, is that they require certain constraints on the behavior of the sellers and on the beliefs of the reporting buyers: i.e. the signals observed by the buyers about the seller's behavior are independently identically distributed, and the set of seller types to which buyers assign positive probability is countable and contains at least 2 elements.

In this paper we address the problem of honest feedback elicitation in a setting in which sellers and buyers are assumed to be rational (i.e. maximize their monetary payoff) and both buyers and sellers have a persistent presence in the market. In Section 2 we prove that persistent presence is a critical assumption for the existence of an incentive-compatible reputation mechanism. Afterwards, we introduce an incentive compatible reputation mechanism and make an analysis of its equilibria. Finally, Section 3 presents some open issues and Section 4 concludes our work.

2 Truthful Feedback

We consider an environment in which the following assumptions hold:

- A rational seller interacts repeatedly with several rational buyers by trading one good of value v_i in each round i. The values $v_i \in (\underline{v}, \overline{v})$ are randomly distributed according to the probability distribution function ϕ [1];
- All transactions generate a fixed profit equal to $(\rho_B + \rho_S)v_i$, where $\rho_S v_i$ is the profit of the seller and $\rho_B v_i$ is the profit of the corresponding buyer. $\rho_B, \rho_S < 1$;
- All buyers are completely trustworthy: i.e. Each buyer first pays the seller and then waits for the seller to ship the good. The seller may defect by not shipping the promised good, and the buyer perfectly perceives the action of the seller;
- There is no independent verification authority in the market, i.e. the behavior of the seller in round i is known only to the seller himself and to the buyer with which he traded in that round;
- The seller cannot refuse the interaction with a specific buyer, and can trade with several buyers in parallel. A buyer can however end the interaction with the seller and choose to buy the goods from a completely trusted seller (e.g. a brick and mortar shop) for an extra cost representing a percentage (θ) of the value of the item bought. Once a buyer decides to terminate a business relationship with the seller, she will never trade again in this market. The seller, however, can always find other buyers to trade with.
- The buyer and the seller discount future revenues by δ_B and δ_S respectively. The discount factors also reflect the probability with which the agents are

[1] Following the same argumentation proposed in [3], this model is valid for settings where the act of accumulating inventory is independent from that of (re)selling it: e.g. a highly dynamic used car dealership.

going to participate to the next transaction. $0 < \delta_S, \delta_B < 1$, and $\delta_S >> \delta_B$ modeling the fact that the seller is likely to have a longer presence in the market than the buyer.
- The buyer and seller interact in a market (possibly a different one for each transaction) capable of charging listing fees and participation taxes.
- At the end of every transaction, both the seller and the buyer are asked to submit a binary report about the seller's behavior: a positive report, $R+$, signals cooperation while a negative report, $R-$, signals defection;

We also assume that in our environment there is a *semantically well defined*, efficient Reputation Mechanism (RM). Reputation is semantically well defined when buyers have exact rules for aggregating feedback into reputation information and for making trust decisions based on that reputation information. These rules determine sellers to assign a value to a reputation report ($R+$ or $R-$), reflecting the influence of that report on future revenues. RM is efficient if the values associated by sellers to reputation reports are such that in any transaction the seller prefers to cooperate rather than defect. If $V(R+, v)$ and $V(R-, v)$ are the values associated by the seller to the positive respectively the negative reputation report generated after a transaction of value v, we have: *V(R+, v) + Payoff(cooperate,v) > V(R−, v) + Payoff(defect,v)*[2]. A simple escrow service or Dellarocas' Goodwill Hunting Mechanism [3] satisfy these properties.

When perfect feedback (i.e. true and accurate) is available, a *well-defined, efficient* RM is enough to make rational sellers cooperate. Unfortunately, perfect feedback cannot be assumed. In the absence of independent verification means, we can only rely on the subjective reports submitted by the agents involved in the transaction; reports which are obviously biased by the strategic interests of the agents.

In the rest of this section we will achieve three things. First, we will draw some limits of feasibility for incentive compatible RMs. We will show that no RM can be incentive compatible when the interaction between the seller and any particular buyer can be modeled by a perfect information finitely repeated game. Second, we describe an incentive-compatible RM that exists within the feasibility bounds. Third, we analyze the equilibria of the described RM and provide an example.

2.1 Limits of Feasibility

From a game theoretic point of view, a perfect information game models a situation in which the players are rational, their rationality is common knowledge and their payoffs are also common knowledge.

Reputation mechanisms cannot exist when the agents have perfect information and the seller is present for a finite number of transactions in the market [10]. It is therefore common practice for RM designers to model sellers by infinite

[2] as an abuse of notation, we will sometimes use $V(R+, v) = V(R+)$ and ignore the fact that the value of a reputation report also depends on the value of the good.

horizon players. However, no restrictions have been imposed so far on the model of the buyers' behavior. This is the problem we address in this section by showing that RMs cannot be incentive compatible when agents have perfect information and any particular buyer is present for a finite number of transactions in the market.

For the environment earlier described, we can prove that:

Theorem 1. *No incentive compatible RM exists in an environment in which the interaction between the seller and a particular buyer can be modeled by a one-shot perfect information game.*

Proof. Consider a single-shot buyer B_i, who trades in round i with the seller having one of the two types: S_1 and S_2. The seller type S_1 cooperates with all buyers, the seller type S_2 cooperates with all buyers but B_i, whom he cheats. Let us assume that there exists an incentive compatible RM. RM will therefore be able to differentiate between the two seller types.

For the rest of the buyers, the behavior types S_1 and S_2 are indistinguishable. Because the behavior of the seller in round i is observed only by B_i (assumption presented in Section 2), the rest of the buyers do not have any information about the truthfulness of buyer B_i's report. Hence, any attempts from the rest of the buyers to bias B_i into telling the truth, even by monetary compensation on the side, would be futile since the "biaser" would have no way of confirming the information of the "biasee". On the other hand, a seller of type S_2 can make any positive (no matter how small) side-payment to buyer B_i in order to convince her to submit a false positive report instead of the true negative one, and B_i, being single-shot, will accept that payment. Therefore RM cannot be incentive-compatible. □

As a direct consequence of Theorem 1, a RM can be incentive-compatible only if it is incentive compatible for every isolated interaction between the seller and a particular buyer. The truth of this statement is evident if we consider that an incentive compatible RM should always be able to distinguish between two seller types that are undistinguishable for all the buyers except one.

As an immediate extension of Theorem 1 we have:

Theorem 2. *No incentive compatible RM exists in an environment in which the interaction between the seller and a particular buyer can be modeled by a perfect information finitely repeated game.*

Proof. Let us denote by N the number of times a buyer trades with the seller, and let us denote by round i_t, $t = 1 \dots N$, the round in which buyer B_i trades for the t^{th} time with the seller. In round i_N the buyer is a one shot buyer, and therefore the result of Theorem 1 applies. Because the outcome of round i_N (in terms of truth reporting) is common knowledge to both the seller and the buyer, it will not influence the outcome of round i_{N-1}, which thus strategically becomes the last interaction. By backward induction, it is not possible to obtain truthful reports in any of the N interactions. □

Fortunately, it still is possible to have an incentive compatible reputation mechanism if (1) either the interaction between the seller and a particular buyer can be modeled by an infinitely repeated game, or (2) agents do not have perfect information. In the remains of this section we will describe a reputation mechanism that supports an incentive compatible equilibrium when agents are perfectly informed. Moreover, in the same spirit as [11], [6] and [14] we show how uncertainty regarding the buyer's type can give birth to the reputation effect and reduce the set of possible equilibria to a more appealing subset.

2.2 The "CONFESS" Mechanism

Every round i, a seller offers for sale a good of value v_i. The market charges the seller a listing fee ε_S, and advertises the good to the buyer. The buyer pays a participation tax ε_B, to the market, and the price v_i to the seller. If the seller cooperates, he ships the good directly to the buyer; otherwise the seller keeps the payment for himself and does not ship the good. After a certain deadline, the transaction is considered as over, and the market starts collecting information about the behavior of the seller. The seller is first required to submit a report. If the seller admits having defected, a negative report ($R-$) is submitted to the RM, the listing fees ε_S and ε_B are returned to the rightful owners, and the protocol is terminated. If, however, the seller claims to have cooperated, the buyer is also asked to provide a report. At this moment, the buyer can report cooperation, report defection, or she can report defection and terminate the interaction with the seller.

If the buyer reports cooperation, a positive reputation report ($R+$) is submitted to the RM, and the listing fees ε_S and ε_B are returned. If the buyer reports defection, both players will be punished as one of them is surely lying: a negative report ($R-$) is submitted to RM, and the listing fees ε_S and ε_B are confiscated. Finally, if the buyer decides to terminate the interaction, a negative report ($R-$) is submitted to RM, and the fees ε_S and ε_B are confiscated.

From a game theoretic point of view, the above described protocol can be modeled by the extensive-form game $G = (N, (A_i), (\succsim_i), T)$, shown in Figure 1. $N = \{S, B\}$ is the set of players, the seller and the buyer respectively, $A_S = \{Cc_S, Cd_S, Dc_S, Dd_S\}$ is the action set of the seller, $A_B = \{c_B, d_B\}$ is the action set of the buyer, $\succsim_S$ is the preference relation of the seller over the set of possible outcomes (g_S is the corresponding payoff function of the seller), $\succsim_B$ is the preference relation of the buyer over the set of possible outcomes (g_B is the corresponding payoff function of the buyer), and T is the player function, or the "turn" function which prescribes which player should make the next move after every possible game history.

The outcome for the buyer is indicated as a single real value representing the buyer's payoff in the current round. The outcome for the seller is indicated as a tuple $(X; P)$, where $X \in \{R+, R-\}$ represents the filed reputation report (positive or negative), and $P \in \mathbb{R}$ is the monetary gain obtained by the seller in the current transaction. The payoff of the seller is defined by simply adding

the monetary gain P with the value of the reputation report: i.e. $g_S(X;P) = V(X) + P$.

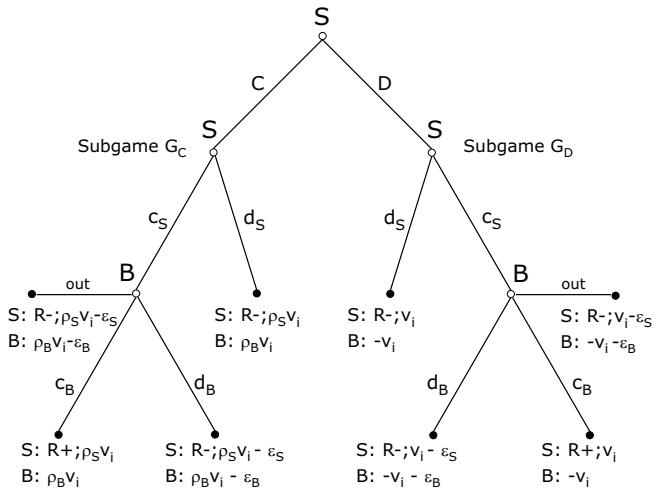


Fig. 1. Game G modeling the one-round interaction protocol.

The repeated transaction between the seller and one buyer can be modeled by an infinite repetition of the stage game G, denoted G^∞, in which the overall payoff for player i is given by the average discounted sum:

$$V_i = (1 - \delta_i) \sum_{\tau=0}^{\infty} \delta_i^\tau g_i^\tau;$$

where δ_i denotes the discount factor of player i, and g_i^τ is the payoff obtained by player i in round τ.

2.3 Equilibrium Analysis

For discounted infinitely repeated games with perfect information, the Folk Theorem [7] guarantees that every enforceable outcome (i.e. feasible and individually rational) can be obtained by a subgame perfect equilibrium (SPE) strategy profile when the discount factors are big enough. The results of this theorem do not apply directly to the game G^∞ because in every round t we allow the buyer to quit the game.

When the buyer terminates an interaction with a seller (chooses *out* in round t), she obtains a continuation payoff equal to:

$$\hat{V}_B^{t+1} = (1 - \delta_B) \sum_{\tau=t+1}^{\infty} \delta_B^{\tau-t-1} v_\tau(\rho_B - \theta);$$

If we denote by $\tilde{v}$ the average value of a transaction, the expected value of $\hat{V}_B^{t+1}$ is:

$$E[\hat{V}_B^{t+1}] = \tilde{v}(\rho_B - \theta);$$

Any SPE strategy profile must give the buyer at least $\hat{V}_B^{t+1}$ after every round t (otherwise the buyer can profitably deviate to *out* in round t). The minimum continuation payoff of the buyer is therefore:

$$\underline{V}_B^t = (1 - \delta_B)(-v_t - \varepsilon_B) + \delta_B \hat{V}_B^{t+1}; \tag{1}$$

A payoff profile $\hat{v} = (\hat{v}_S, \hat{v}_B)$ dominates another payoff profile $v = (v_S, v_B)$ if it is better for at least one of the players and not worse for any of the players: i.e. there is $i \in \{S, B\}$ such that $\hat{v}_i > v_i$ and for all $j \in \{S, B\} \setminus i$, $\hat{v}_j \geq v_j$.

We restrict our attention to SPE strategies of G^∞ which are not dominated. A SPE strategy s is not dominated if there is no other SPE strategy $\hat{s}$ such that the the payoff profile generated by $\hat{s}$ dominates the payoff profile generated by s in G^∞. The intuition behind this assumption is that no player will choose to play a SPE strategy as long as there is another SPE strategy which can bring him a higher payoff while not decreasing the payoff of the opponent.

The above restriction limits the set of SPE strategies to the ones generating an equilibrium path containing a mixture of the action profiles (Cc_S, c_B) and (Dc_S, c_B).

Lemma 1. *All not dominated SPE strategies prescribe only the action profiles* (Cc_S, c_B) *and* (Dc_S, c_B) *on the equilibrium path in* G^∞.

Proof. See Jurca and Faltings, [9], Lemma 1 □

Let s be a mixed strategy profile such that with probability p the players play (Dc_S, c_B) and with probability $(1-p)$ the players play (Cc_S, c_B). The expected continuation payoff of the buyer is:

$$\begin{aligned} E[V_B^{t+1}] &= E\left[(1 - \delta_B) \sum_{\tau=t}^{\infty} \delta_B^{\tau - t}[p(-v_\tau) + (1-p)\rho_B v_\tau]\right]; \\ &= \tilde{v}(\rho_B - p - \rho_B p); \end{aligned} \tag{2}$$

When playing in round t, the buyer knows which of the action profiles (Cc_S, c_B) or (Dc_S, c_B) are prescribed by the strategy s, and therefore the continuation payoff of the buyer is:

$$\begin{aligned} V_B^t|_{(Cc_S, c_B)} &= (1 - \delta_B)\rho_B v_t + \delta_B V_B^{t+1}; \\ V_B^t|_{(Dc_S, c_B)} &= (1 - \delta_B)(-v_t) + \delta_B V_B^{t+1}; \end{aligned} \tag{3}$$

depending on what s prescribes for round t. Since both $V_B^t|_{(Cc_S, c_B)}$ and $V_B^t|_{(Dc_S, c_B)}$ have to be greater or equal to $\underline{V}_B^t$, the maximum value of p is:

$$p \leq \overline{p} = \frac{(1 - \delta_B)\varepsilon_B + \delta_B \tilde{v}\theta}{\delta_B \tilde{v}(1 + \rho_B)}; \tag{4}$$

The upper bound on p limits the maximum attainable payoff, $\overline{V}_S$ of the seller in G^∞:

$$\overline{V}_S^t = (1-\delta_S)\sum_{\tau=t}^{\infty}\delta^{\tau-t}[\overline{p}v_\tau + (1-\overline{p})\rho_S v_\tau + V(R+)];$$

which has an expected value: $E[\overline{V}_S^t] = V(R+) + \overline{p}\tilde{v}(1-\rho_S) + \tilde{v}\rho_S$. By replacing (4) we obtain:

$$\overline{V}_S^t = V(R+) + \tilde{v}\rho_S + \tilde{v}(1-\rho_S)\frac{(1-\delta_B)\varepsilon_B + \delta_B\tilde{v}\theta}{\delta_B\tilde{v}(1+\rho_B)};$$

For any $p \in [0, \overline{p}]$ the strategy s can be made a SPE of G^∞ by adding minimax threats (see Fudenberg and Maskin [7] Therem 1 for how the strategy can be built). Let us observe that when $p = 0$ no false feedback is recorded by the RM, and every transaction is cooperative. "CONFESS" has therefore one incentive-compatible, efficient SPE point. Unfortunately, this equilibrium is not unique, and we can only guarantee that the maximum percentage of false reports accepted by our mechanism is $\overline{p}$.

Incomplete Information. Following the ideas from [10], [6] and [14] we can use imperfect information in order to limit the set of SPE strategies to a more desirable subset (i.e. consisting of those strategies which generate mainly true reputation reports and outcomes as close as possible to the socially efficient one).

Let us consider a perturbation of the complete information repeated game G^∞ such that in period 0 (before the first round of the game is played) the *type* of the buyer is drawn by nature out of the set $\Omega = \{\omega_0, \omega^*\}$ according to the probability measure μ. The buyer's payoff now additionally depends on her type, such that the ω_0 type buyer (or the *normal type* buyer) has the payoffs presented in Figure 1, while the ω^* type buyer (or the *commitment type* buyer) always prefers to report the truth. We say that in the perturbed game $G^\infty(\mu)$ the seller has incomplete (or imperfect) information because he is not sure about the true type of the buyer.

We prove that in $G^\infty(\mu)$ there is a finite upper bound, k_S, on the number of times a rational seller is willing to play Dc_S, given that he always observes the commitment strategy being played by the buyer.

The intuition behind this result is the following. The seller's best response to the commitment type buyer is to always cooperate and report cooperation, i.e. (Cc_S), which gives the commitment type buyer her maximum attainable payoff in $G^\infty(\mu)$, corresponding to the socially efficient outcome. The seller however would be better off by playing against the normal buyer. As we have seen above, against the normal type buyer, the seller can get more than the cooperative outcome by randomizing between the (Cc_S, c_B) and (Dc_S, c_B) action profiles.

A normal type buyer can be distinguished from a commitment type buyer only if the seller plays Dc_S. In this situation, the normal buyer prefers to play c_B, while the commitment buyer prefers to play d_B. The normal buyer could

however simulate the strategy of a commitment buyer in order to obtain the payoff of the latter (i.e. the cooperative outcome).

Because the cooperative strategy involves a loss for the seller (i.e. the potential loss of not being able to get the higher payoff that could be obtained against the normal buyer) the seller should not become "easily" convinced that he is playing against a commitment type buyer. The question is therefore, how long should the seller try to determine the true type of the buyer. Because every outcome (Dc_S, c_B) (i.e. the seller tests the type of the buyer and the buyer plays the commitment strategy) generates a loss for the seller, and because the seller cannot wait infinitely for future payoffs (the seller's discount factor is less than 1) it follows that at some point, if the seller always observes the commitment strategy being played by the buyer, he must give up trying to test the true type of the buyer, and accept playing a best response to the commitment type buyer.

Theorem 3. *If:*

1. *the seller has incomplete information in G^∞,*
2. *the seller assigns positive probability to the prior beliefs that the buyer is a "commitment" type and a "normal" type. i.e. $\mu(\omega_0) > 0$, $\mu_0^* = \mu(\omega^*) > 0$ and $\mu(\omega_0) + \mu(\omega^*) = 1$;*

Then there is a finite upper bound k_S on the number of times the seller plays Dc_S in G^∞:

$$k_S = \left\lceil \frac{\ln(\mu_0^*)}{\ln\left(\frac{(1-\delta_S)\overline{v}(1-\rho_S)+\delta_S\Phi}{(1-\delta_S)[\overline{v}(1-\rho_S)+\epsilon+\varepsilon_S]+\delta_S\Phi}\right)} \right\rceil$$

where δ_B, δ_S are the discount factors of the buyer and seller, ρ_S, ρ_B are their profit margins, $\varepsilon_B, \varepsilon_S$ are the lying fines imposed by the mechanism, $\overline{v}$ is the maximum value of a transaction, $\tilde{v}$ is the expected value of a transaction, θ is the additional fraction of the price a buyer has to pay when buying from completely trustworthy sellers, $\epsilon = g_S(R+, \rho_S v_i) - g_S(R-, v_i)$ is the loss of the seller caused by receiving a negative reputation report instead of a positive one, and:

$$\Phi = \tilde{v}(1-\rho_S)\frac{(1-\delta_B)\varepsilon_B + \delta_B\tilde{v}\theta}{\delta_B\tilde{v}(1+\rho_B)};$$

Proof. See Jurca and Faltings, [9], Theorem 1 □

The lower bound k_S restricts the set of possible equilibrium payoffs of the normal type buyer in $G^\infty(\mu)$. If a rational buyer mimics the commitment type buyer, she obtains in the worst case $\underline{V'}_B^t$; the outcome (Dc_S, c_B) in the first k_S rounds, followed by an infinite number of cooperative outcomes.

$$\underline{V'}_B^t = (1-\delta_B)\Bigg[(-v_t - \varepsilon_B) + \delta_B \sum_{\tau=t+1}^{t+k_S-1} \delta_B^{\tau-t-1}(-v_\tau - \varepsilon_B) + \delta_B^{k_S} \sum_{\tau=t+k_S}^{\infty} \delta_B^{\tau-t-k_S}\rho_B v_\tau\Bigg];$$

Any equilibrium strategy in $G^{\infty}(\mu)$ must guarantee the normal type buyer at least $\underline{V'}_B^t$. Let us reconsider the strategy s from the perfect information game G^{∞} according to which the players play (Dc_S, c_B) with probability p and (Cc_S, c_B) with probability $1-p$. By imposing that both $V_B^t|_{(Cc_S,c_B)}$ and $V_B^t|_{(Dc_S,c_B)}$ (Equation (3)) be greater or equal to $\underline{V'}_B^t$, the maximum value of p is:

$$p \leq \overline{p'} = \frac{(1-\delta_B)\varepsilon_B + (\delta_B - \delta_B^{k_S})(\tilde{v} + \varepsilon_B + \tilde{v}\rho_B)}{\delta_B \tilde{v}(1+\rho_B)}; \tag{5}$$

However, the constraints on p presented in Equation (4) remain valid, and therefore $p \leq \min(\overline{p}, \overline{p'})$.

Particular importance has the case in which $k_S = 1$. $\overline{p'}$ becomes:

$$\overline{p'} = \frac{(1-\delta_B)\varepsilon_B}{\delta_B \tilde{v}(1+\rho_B)}; \tag{6}$$

and as ε_B can be any positive value, $\overline{p'}$ will in the limit approach 0. In this situation, the reputation mechanism will receive false reputation reports with vanishing probability.

The result of Theorem 3 has to be interpreted as a worst case scenario. In real markets, sellers that already have a small predisposition to cooperate will defect fewer times. Moreover, the mechanism is self enforcing, in the sense that the more buyers act as commitment types, the higher will be the prior beliefs of the sellers that buyers will report truthfully, and therefore the easier it will be for the buyers to act as truthful reporters.

The following properties are also straightforward to derive as a direct consequence of Theorem 3:

Property 1. The mechanism is bounded socially efficient.

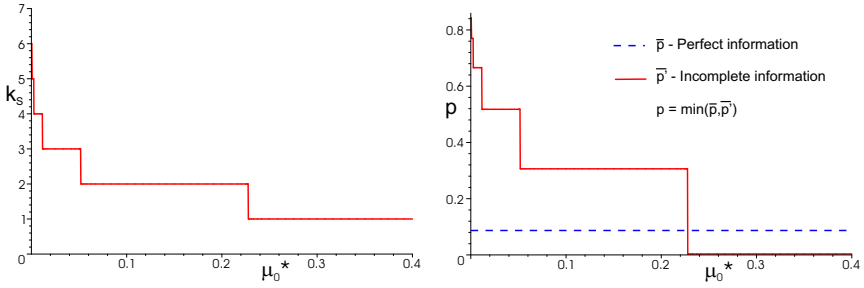
Proof. Because of the lost exchange, outcome (Dc_S, c_B) generates a cumulated social loss of $(\rho_S+\rho_B)v_i$ every time it occurs. The perfect information equilibrium involves a possibly infinite number of rounds in which (Dc_S, c_B) is played. By limiting the number of times the seller is playing action D, we also limit to a finite number (i.e. k_S) the rounds in which the exchange does not occur. The social loss is therefore bounded above by $k_S(\rho_S + \rho_B)\overline{v}$. □

Property 2. The mechanism is weakly budget balanced

Proof. The net payment to the mechanism is non-negative as every time there is a disagreement concerning the two reputation reports, the center gets $\varepsilon_B+\varepsilon_S$. By introducing supplementary service fees, the mechanism can be easily transformed into one that yields profit to the market. □

Numerical Example. Let us consider the example of a hotel who charges for a room a fixed price of $v = 140$ dollars a night. The profit margin of the hotel is $\rho_S = 0.95$ while the profit margin of the client is $\rho_B = 0.2$. The customer returns to the same hotel once a year with probability $\delta_B = 0.7$, and after each night spent in the hotel she is required to submit a binary reputation report about whether or not the hotel has kept its promise (in terms of a Service Level Agreement). The customer also has the option to go to another hotel which costs an additional 14 dollars a night ($\theta = 0.1$). The hotel discounts future revenues with $\delta_S = 0.95$.

We assume that the reputation of the hotel directly affects its occupancy (and future revenues) such that any time a hotel cheats and correctly receives a negative reputation report, it looses (in terms of future revenues) $\epsilon = 25$ dollars. When the fines ε_S and ε_B equal to 1 respectively 20 dollars, Figure 2(a) plots the value of the upper bound k_S for different values of the prior probability μ_0^*. For the same values of μ_0^*, Figure 2(b) plots the maximum value of the probability with which "CONFESS" will accept false reputation reports. When $\mu_0^* > 0.25$ the hotel will cheat at most once on a customer, and the probability of receiving a false reputation report is smaller than 0.3%.



(a) The upper limit k_S depending on μ_0^* (b) The probability of receiving a false report depending on μ_0^*

Fig. 2. Numerical Example

3 Open Issues

Further benefits can be obtained if the buyers' reputation as honest reporters is shared within the market. A buyer that has once built a reputation for truthfully reporting the seller's behavior will benefit from cooperative trade during her entire lifetime, without having to convince each seller separately. Therefore the upper bound on the loss a buyer has to withstand in order to convince a seller that she is a commitment type, becomes an upper bound on the total loss a buyer has to withstand during her entire lifetime in the market. How to efficiently share the reputation of buyers within the market remains an open issue.

Correlated with this idea is the observation that buyers that use our mechanism are motivated to keep their identity. In generalized markets in which agents are encouraged to play both roles (e.g. a peer-2-peer file sharing market in which the fact that an agent acts only as "seller" can be interpreted as a strong indication of "double identity" with the intention of cheating) our mechanism also solves the problem signaled in [5] related to the ease with which agents can change their online identity. The price to pay for the new identity is the loss due to building a reputation as a honest reporter when acting as a buyer.

The mechanism can be criticized for being centralized. The market acts as a central authority by collecting listing fees from the seller and the buyer, by asking the reputation reports at the end of each transaction, and by reasoning about the outcome of the transaction. However, as the mechanism does not require any information to be transmitted from one round to another (the seller stores the reputation of the buyer) we could have the same seller and buyer interact in multiple markets (decentralized system) without having to rely on one single centralized institution.

One direction of future research is to study the behavior of the above mechanism when there is two-sided incomplete information: i.e. the buyer is also uncertain about the type of the seller. A seller type of particular importance would be the "greedy" seller type who always likes to keep the partner buyer to her minimum continuation payoff. In this situation we expect to be able to find an upper bound k_B on the number of rounds in which a rational buyer would be willing to test the true type of the seller. The condition $k_S < k_B$ would impose the constraints on the parameters of the system for which the reputation effect will work in the favor of the buyer: i.e. the seller will give up first the "psychological" war and revert to a cooperative equilibrium.

A somehow related problem is the robustness to mistakes, or imperfect monitoring of the opponent's actions. A seller's defection by mistake in a situation in which it was not rational for a seller to defect will be interpreted by the buyer as evidence of the seller's irrational behavior.

Last, but not least, we plan to adapt this truthful reporting mechanism for reputation mechanism that affect the value of future transactions. For such mechanisms the repeated interaction between a buyer and seller is much more complicated to model. A negative report submitted by a buyer at time t might lead to more beneficial trade for that buyer in the future (since the negative reputation report will attract a decrease in the price of future sold goods). Making it rational for the buyer to submit the true report involves a detailed understanding of the underlying reputation mechanism, the solution being most likely application dependent.

4 Conclusions

In this paper we formally prove that no binary reputation mechanism can be incentive compatible when the agents are rational, have game-theoretic perfect information and the trusting agent (i.e. the buyer) interacts a finite number of

times with the trusted agent (i.e. the seller). Moreover, we describe a truthful feedback elicitation mechanism ("CONFESS") for two long-run rational buyer and seller and give an intuitive presentation of how incentive compatibility can exist as an equilibrium. When the seller has imperfect information, the performance of our mechanism is greatly improved and we have been able to derive an upper bound on the percentage of false reports that are accepted by the mechanism. The mechanism we have presented does not require the presence of an independent verification authority, can be easily decentralized and accepts transactions of different values.

References

1. A. Birk. Learning to Trust. In R. Falcone, M. Singh, and Y.-H. Tan, editors, *Trust in Cyber-societies*, volume LNAI 2246, pages 133–144. Springer-Verlag, Berlin Heidelberg, 2001.
2. A. Biswas, S. Sen, and S. Debnath. Limiting Deception in a Group of Social Agents. *Applied Artificial Intelligence*, 14:785–797, 2000.
3. C. Dellarocas. Goodwill Hunting: An Economically Efficient Online Feedback. In J. Padget and et al., editors, *Agent-Mediated Electronic Commerce IV. Designing Mechanisms and Systems*, volume LNCS 2531, pages 238–252. Springer Verlag, 2002.
4. C. Dellarocas. Efficiency and Robustness of Binary Feedback Mechanisms in Trading Environments with Moral Hazard. MIT Sloan Working Paper #4297-03, 2003.
5. E. Friedman and P. Resnick. The Social Cost of Cheap Pseudonyms. *Journal of Economics and Management Strategy*, 10(2):173–199, 2001.
6. D. Fudenberg and D. Levine. Reputation and Equilibrium Selection in Games with a Patient Player. *Econometrica*, 57:759–778, 1989.
7. D. Fudenberg and E. Maskin. The Folk Theorem in Repeated Games with Discounting or Incomplete Information. *Econometrica*, 54(3):533–554, 1989.
8. R. Jurca and B. Faltings. An Incentive-Compatible Reputation Mechanism. In *Proceedings of the IEEE Conference on E-Commerce*, Newport Beach, CA, USA, 2003.
9. R. Jurca and B. Faltings. Truthful reputation information in electronic markets without independent verification. Technical Report ID: IC/2004/08, EPFL, http://ic2.epfl.ch/publications, 2004.
10. D. M. Kreps, P. Milgrom, J. Roberts, and R. Wilson. Rational Cooperation in the Finitely Repeated Pisoner's Dilemma. *Journal of Economic Theory*, 27:245–252, 1982.
11. D. M. Kreps and R. Wilson. Reputation and Imperfect Information. *Journal of Economic Theory*, 27:253–279, 1982.
12. P. Milgrom and J. Roberts. Predation, Reputation and Entry Deterrence. *J. Econ. Theory*, 27:280–312, 1982.
13. N. Miller, P. Resnick, and R. Zeckhauser. Eliciting Honest Feedback in Electronic Markets. Working Paper, 2003.
14. K. M. Schmidt. Reputation and Equilibrium Characterization in Repeated Games with Conflicting Interests. *Econometrica*, 61:325–351, 1993.
15. R. Selten. The Chain-Store Paradox. *Theory and Decision*, 9:127–159, 1978.
16. B. Yu and M. Singh. An Evidential Model of Distributed Reputation Management. In *Proceedings of the AAMAS*, Bologna, Italy, 2002.