Eliminating Undesired Equilibrium Points from Incentive Compatible Reputation Mechanisms

Radu Jurca and Boi Faltings

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Artificial Intelligence Laboratory
CH-1015 Lausanne, Switzerland
{radu.jurca, boi.faltings}@epfl.ch

Abstract. Choosing the best out of an increasing number of options requires reliable and accurate information. As our time and resources are limited, we commonly use the experience of others in order to take decisions. Reputation mechanisms aggregate in a formal way the feedback collected from peers and compute the "reputation" of products, services, or providers. They enjoy huge success and are believed to be the key of the agent mediated commerce of tomorrow.

Obtaining honest feedback from self-interested agents is not a trivial problem. Mechanisms based on side-payments can be conceived such that honest reporting becomes rational (i.e. Nash equilibrium). Unfortunately, for every incentive-compatible Nash equilibrium there seems to also be a dishonest Nash equilibrium strategy that sometimes is more attractive. In this paper we analyze two incentive-compatible reputation mechanisms and investigate how undesired equilibrium points can be eliminated by using trusted (i.e. true) reports.

1 Introduction

In a world that offers an ever increasing number of options, having the information to make the right choices becomes of vital importance. The internet, and advances in communication technologies have made raw data readily available anytime, anywhere. It is trivial today to get extensive technical descriptions of computers, TV-sets, or any other products you might want to buy. Comparative charts between similar products of different brands are a few more mouse-clicks away, however, making sense of all this data still requires intensive human effort. The usefulness of increasing amounts of data is thus limited by the processing capacity of the human brain and by the available physical time. Moreover, even the most detailed technical charts cannot answer crucial questions like: "*How useful or reliable is this product?*" or "*How well does the manufacturer fulfill its promises?*" in terms of product quality and customer support.

It is therefore common practice to draw on the experience of peers in order to make decisions. Aggregated feedback coming from previous users constitutes the *reputation* of a product, service or manufacturer. Reputation information accounts for the data we cannot directly observe before the purchase (e.g. reliability of the product/manufacturer, overall quality, hidden aspects, etc) and also allows to reuse the effort spent by previous buyers in collecting and analyzing available data. Moreover, by making reputation information machine-readable (through clear syntax and semantics) it can also be used by software agents in the automated e-market of tomorrow.

Reputation Mechanisms (RM) are responsible for collecting and aggregating feedback. They enjoy huge success and have become a mandatory component of every online market. Early implementers (e.g. eBay[1] or Amazon[2]) owe part of their success to such “feedback forums” which make it very easy for subsequent buyers to assess the trustworthiness/quality of a seller or book. Studies show that buyers seriously take into account the reputation of the seller when placing their bids in online auctions [9] and that despite the incentive to free ride, feedback is provided in more than half of the transactions on eBay [16].

Obtaining honest feedback from agents is not a trivial problem. Rational agents report according to their selfish interests. For example, disclosing the genuine positive experience with a scarce service or product is never rational. The resulting increase in reputation will attract more consumers and therefore less availability for the reporter in the future: e.g. smart parents do not disclose the name of their favorite babysitter [5].

As a consequence, RM have to provide the right incentives in order to convince rational agents to report the truth (incentive-compatibility). One way to do it is to pay for feedback reports according to their estimated truthfulness. Since verification authorities are usually not available, the truthfulness of a report is assessed by comparing it with reports coming from other agents about the same service or service provider.[3] It turns out that if there is sufficient correlation between reports, there exist payment rules that make truthful reporting a Nash Equilibrium (i.e. rational agents report the truth given that all other agents report the truth).

The basic idea behind this class of incentive-compatible reputation mechanisms can be grasped by considering the example of a webservice that is sequentially invoked by a set of users. Despite the best intention of the owning company, there is some fixed probability that the service invocation fails. Feedback reports about such a webservice are binary values, (i.e. the invocation failed or not) and the reputation of the webservice will estimate its failure rate. Since all feedback reports refer to the same random event, there is some degree of correlation between them. Future reports (assuming that they are true) can give information about the honesty of the present report. Side payments considering this information, can make it rational for agents to report the truth. Consequently, the

[1] www.ebay.com

[2] www.amazon.com

[3] The term “service” is used generically and can stand for any product (e.g. book) or service (e.g. web-hosting). Likewise, the “service provider” will stand for the entity providing the service (e.g. the author of the book, the ISP providing the web-hosting service)

system has a Nash equilibrium incentive compatible equilibrium point. [15] and [10] describe such concrete mechanisms.

Unfortunately, such mechanisms typically have multiple Nash equilibrium points. Taking the example from the previous paragraph, always reporting positive feedback is also a Nash Equilibrium strategy. For example, 99.1% of the feedback submitted on eBay is positive [7]. From a game theoretic point of view this might be regarded as evidence for the existence of a cooperative Nash equilibrium of the repeated trading game [8]. However, it can also be the manifestation of a Nash equilibrium reporting strategy in which every agent reports positive feedback.

Moreover, the incentive compatible equilibrium payoff is often dominated by some other Nash equilibrium payoff. Whenever an agent truthfully reports negative feedback, it risks retaliation from the subject of the report. On eBay, a negative report submitted by a buyer about a transaction is often followed by a negative report submitted by the seller about the buyer. Therefore, it is more profitable to adopt an equilibrium point where you only report positive reports rather than true reports. The mechanisms described in [15] and [10] suffer from the same drawback.

Finally, the existence of multiple equilibrium points is a serious impediment for the engineering of reputation mechanisms in real application. The behavior of such mechanisms can be very chaotic, and there is no guarantee that the agents will choose exactly the incentive compatible strategy over one of the many other dishonest equilibrium strategies.

Trusted reports (verifiable reports coming from independent authorities) can be used to eliminate the undesired Nash equilibrium points. In the extreme case in which the feedback provided by agents is compared only against trusted (i.e. true) reports, the incentive compatible Nash equilibrium becomes unique. However, in most cases it is enough to have trusted reports for comparison, only with a certain probability. Since trusted reports are expensive, it is interesting to determine how small this probability can be.

In this paper we investigate the influence of trusted reports on the set of equilibrium points of the reputation mechanisms described in [15] and [10]. While the specific results of this paper are only applicable to the two mechanisms mentioned above, this paper introduces a general methodology for studying the influence of trusted reports on the equilibrium points of a reputation mechanism. Section 2 introduces the setting and explains the functioning of the two mechanisms. Section 3 analysis the Nash equilibrium points of the mechanisms and analytically shows how trusted reports can be used to eliminate the undesired equilibria. Numerical results are presented and interpreted in Section 4, followed by related work and a conclusion.

2 The Setting

In this section we present in detail the two reputation mechanisms that will be analyzed in the rest of the paper. The two mechanisms have been named after the initials of their authors: i.e. The MRZ mechanism denotes the mechanism described by Miller, Resnick and Zeckhouser in [15] and the JF mechanism denotes the mechanism described by Jurca and Faltings in [10].

2.1 The MRZ Incentive Compatible Reputation Mechanism

In [15], Miller et al. consider that a number of agents sequentially experience the same service whose *type*[4] is drawn from a set of possible types T.[5]

The real type of the service does not change during the experiment, and is not known by the agents. However, after every interaction, the agent receives one signal s (from a possible set of signals S of cardinality M) about the type of the service. For a certain product type $t \in T$, the signals perceived by the agents are independently identically distributed such that the signal s_i is observed with probability $f(s_i|t)$ for all $s_i \in S$. $\sum_{s_i \in S} f(s_i|t) = 1$ for all $t \in T$.

After every interaction, the participating agent is asked to submit feedback about the signal she has observed. The reputation mechanism collects the reports, and updates the reputation of the product. Reputation information consists of the probability distribution over the possible types of the service. Let p be the current belief of the reputation mechanism (and therefore of all agents that can access the reputation information) about the probability distribution over types of the service. $p(t)$ is the probability assigned by the current belief to the fact that the service is of type t, and $\sum_{t \in T} p(t) = 1$. When the reputation mechanisms receives a report $r \in S$, the belief p is updated using Bayes' Law:

$$p(t|r) = \frac{f(r|t) \cdot p(t)}{Pr[r]}$$

where $Pr[r] = \sum_{t \in T} f(r|t) \cdot p(t)$ is the probability of observing the signal r.

Each feedback report is compared against another future report. Let $r \in S$ be the report submitted by agent a and let r_r be the future report submitted by agent a_r which serves to assess the honesty of r. The agent a_r is called the *rater* of the agent a since the report r_r is used to "rate" the report r. Typically, the next report is used to evaluate the present one. Miller et al. show that if agent a is paid according to the scoring rule $R(r_r, r)$, she will honestly report her observation given that a_r also honestly reports his observation. The three best known scoring rules are:

1. Quadratic Scoring Rule: $R(r_r, r) = 2Pr[r_r|r] - \sum_{s_h \in S} Pr[s_h|r]^2$

[4] The type of a service defines the totality of relevant characteristics of that service. For example, quality, and possibly other attributes define the type of a product.

[5] The set of possible types is the combination of all values of the attributes which define the type. While this definition generates an infinite-size set of types, in most practical situations, approximations make the set of possible types countable. For example, the set of possible types could have only two elements: *good* and *bad*. This implies that there is common understanding among the agents in the environment that the service can be exhaustively classified as good or bad: i.e. no other information about the service is relevant for the decision taken by the agents in that environment.

2. Spherical Scoring Rule: $R(r_r, r) = \frac{Pr[r_r|r]}{\left(\sum_{s_h \in S} Pr[s_h|r]^2\right)^{1/2}}$
3. Logarithmic Scoring Rule: $R(r_r, r) = \ln Pr[r_r|r]$

where $Pr[s_h|r] = \sum_{t \in T} f(s_h|t) \cdot p(t|r)$ is the posterior probability that the signal s_h will be observed, as known by the reputation mechanism immediately after r has been reported.

To illustrate how this mechanism works, let us consider that a service can have two possible types, i.e., *good* (G) or *bad* (B). Buyers can observe two signals, $+$ or $-$ such that the distribution of signals conditional of the type of the service is: $f(+|G) = 0.9$, $f(-|G) = 0.1$, $f(+|B) = 0.15$, $f(-|B) = 0.85$. Let us assume that the current belief about the type of the service assigns probability 0.4 to the service being *good* and 0.6 to the service being *bad.* i.e. $p(G) = 0.4$ and $p(B) = 0.6$.

Let us assume that the agent a has the next interaction with the service, and that she has observed a $+$. Figure 1 shows how the side-payments are computed, and how beliefs are updated if a reports $+$ or $-$.

Given that a_r reports the truth, a maximizes her expected payoff by also reporting the truth. The same would be true if a had observed a $-$ rather than a $+$. Miller et al. show that in every situation (for every signal observed, for every belief about the service, and for all generic distributions of signals conditional on types) it is better for a to report the truth, given that a_r also reports the truth. Honest reporting is therefore a Nash equilibrium.

However, there are also other Nash equilibrium strategies that are not incentive-compatible. Examples of such strategies will be presented in Section 3.1.

2.2 The JF Incentive-Compatible Reputation Mechanism

The MRZ mechanism can be easily adapted to a variety of contexts. However, it assumes (1) common knowledge about the distribution of signals conditional on types and (2) lack of private information.

Since the MRZ mechanism is based on side-payments which are computed using scoring rules, the agents always have the incentive to approximate as good as possible the signal that will be observed by the rater. When the two assumption above are satisfied, the incentive to provide the best approximation for the signal received by the rater coincides with honestly reporting the observed signal. However, when the reporting agent and the reputation mechanism have different views of the world (i.e. different beliefs about the service), the agent can manipulate her report depending on her private beliefs (about the service and about the beliefs of the reputation mechanism).

The JF mechanism eliminates this drawback at the expense of limiting the contexts in which the incentive-compatible property holds. The model used by Jurca and Faltings in [10] is that of a service having a “dynamic type”. The signals perceived by the agents do not only depend on the type of the service, but also on temporary information. The model adopted for the probability distribution of signals is that of a Markov chain of variable length, and the possible set of signals consist of only two values $+$ and $-$.

<table>
<tr><th>a reports $+$</th><th>a reports $-$</th></tr>
<tr><td colspan="2">Beliefs of a regarding the posterior distribution over types</td></tr>
<tr><td colspan="2">$p(G|+) = \frac{f(+|G) \cdot p(G)}{f(+|G) \cdot p(G) + f(+|B) \cdot p(B)} = 0.8;$
$p(B|+) = 1 - p(G|+) = 0.2$</td></tr>
<tr><td colspan="2">Beliefs of the Reputation Mechanism regarding the posterior distribution over types</td></tr>
<tr><td>$p(G|+) = \frac{f(+|G) \cdot p(G)}{f(+|G) \cdot p(G) + f(+|B) \cdot p(B)} = 0.8;$
$p(B|+) = 1 - p(G|+) = 0.2$</td><td>$p(G|-) = \frac{f(-|G) \cdot p(G)}{f(-|G) \cdot p(G) + f(-|B) \cdot p(B)} = 0.07;$
$p(B|-) = 1 - p(G|-) = 0.93$</td></tr>
<tr><td colspan="2">Beliefs of a regarding the distribution of signals received by a_r</td></tr>
<tr><td colspan="2">$Pr[+|+] = f(+|G) \cdot p(G|+) + f(+|B) \cdot p(B|+) = 0.75$
$Pr[-|+] = 1 - Pr[+|+] = 0.25$</td></tr>
<tr><td colspan="2">Beliefs of the Reputation Mechanism regarding the distribution of signals received by a_r</td></tr>
<tr><td>$Pr[+|+] = f(+|G) \cdot p(G|+) + f(+|B) \cdot p(B|+) = 0.75$
$Pr[-|+] = 1 - Pr[+|+] = 0.25$</td><td>$Pr[+|-] = f(+|G) \cdot p(G|-) + f(+|B) \cdot p(B|-) = 0.2$
$Pr[-|-] = 1 - Pr[+|-] = 0.8$</td></tr>
<tr><td colspan="2">Payment made to a (Using spherical scoring rule)</td></tr>
<tr><td>$R(+,+) = \frac{Pr[+|+]}{\sqrt{Pr[+|+]^2 + Pr[-|+]^2}} = 0.95$ if $r_r = +$
$R(-,+) = \frac{Pr[-|+]}{\sqrt{Pr[+|+]^2 + Pr[-|+]^2}} = 0.32$ if $r_r = -$.</td><td>$R(+,-) = \frac{Pr[+|-]}{\sqrt{Pr[+|-]^2 + Pr[-|-]^2}} = 0.24$ if $r_r = +$
$R(-,-) = \frac{Pr[-|-]}{\sqrt{Pr[+|-]^2 + Pr[-|-]^2}} = 0.97$ if $r_r = -$.</td></tr>
<tr><td colspan="2">Expected payment to a</td></tr>
<tr><td>$E_{s_j \in \{+,-\}}[R(s_j|+)] = Pr[+|+] \cdot R(+,+) + Pr[-|+] \cdot R(-,+) = 0.79$</td><td>$E_{s_j \in \{+,-\}}[R(s_j|-)] = Pr[+|+] \cdot R(+,-) + Pr[-|+] \cdot R(-,-) = 0.42$</td></tr>
</table>

Fig. 1. Updating of beliefs, and computation of side payments according to the MRZ mechanism, given that a has observed a $+$.

The side-payment for reports follows a very simple rule, and does not depend on the beliefs of the agent or of the reputation mechanism. A report is paid only if the next report submitted about the same service has the same value. The amount of the payment is dynamically scaled such that the whole mechanism is budget-balanced.

The Markov model for the observable signals is very appropriate for services offered by software agents. Let us recall the service used in section 2.1, and let us also consider that the service is provided by a software agent (i.e. a webservice). One possibility is to consider that the webservice is always providing the same service (*good* or *bad*) and that the agents perceive the signals $+$ and $-$ with the probabilities: $f(+|G), f(-|G), f(+|B)$ and $f(-|B)$. However, if the *good* and *bad* types are interpreted as successful, respectively defective service (and the $+$ and $-$ signals are interpreted as satisfactory, respectively unsatisfactory answers, perceived with some inherent noise) it is more realistic to assume that failures are correlated. Intuitively, the failure of the present invocation is an indication of exceptional conditions (hardware failure, blocks in the software, overload, etc) and therefore is likely to influence the result of the next invocation. For example, a failure of the present invocation due to hardware problems indicates a big probability of failure for the next invocation as well. On the contrary, present failure due to an overload might indicate a bigger probability of success for the next invocation.

While the MRZ mechanism can easily be adapted for Markov models of behavior, it requires that the model be common knowledge among the agents: i.e. all agents must agree on the length of the model and on the fact that there is a unique set of parameters characterizing that model. By having side-payments that do not depend on the beliefs of the agents, the JF mechanism allows the agents to have any private beliefs about the model of the webservice, as long as these beliefs satisfy some general constraints. Of course, the freedom of having private beliefs is paid by the constraints which must be satisfied, that limit the contexts in which incentive-compatibility is guaranteed.

3 Eliminating Undesired Equilibrium Points

3.1 Nash Equilibria of the original mechanisms

Formally, a pure reporting strategy of an agent a is a mapping $\sigma : S \to S$ such that $\sigma(s_i) \in S$ is the signal reported by a when she observes the signal s_i. Similarly, a mixed reporting strategy is a mapping from the set of signals S to the set of all probabilistic combinations of signals from S. $\sigma(s_i) = \sum_{j \in S} \alpha_j^i s_j$ denotes that s_j is reported with probability α_j^i given that the signal observed by a was s_i. $\sum_{j=1}^{M} \alpha_j^i = 1$ for all $i \in \{1, \dots, M\}$.

The incentive compatible strategy is denoted by σ^* such that $\sigma^*(s_i) = s_i$ for all $s_i \in S$. By an abuse of notation we also use s_j to denote the "constant" reporting strategy: $s_j(s_i) = s_j$ for all $s_i \in S$.

Given that a uses reporting strategy σ and that the rater of a (i.e. a_r) uses reporting strategy $\sigma' = (\beta_j^i)$, the expected payment of a when observing the signal s_i is:

$$E[\sigma, \sigma', s_i] = \sum_{j=1}^{M} \alpha_j^i \left(\sum_{k=1}^{M} Pr[s_k | s_i] \cdot \left(\sum_{l=1}^{M} \beta_l^k \cdot R(s_l, s_j) \right) \right);$$

where:

- $Pr[s_k|s_i]$ is the probability that the rater observes the signal s_k given that a has observed s_i,
- the function $R(s_l, s_j)$ gives the payment made by the reputation mechanism to a if a reports the signal s_j and a_r reports the signal s_l.

For the MRZ mechanism the function $R(s_l, s_j)$ is defined by one of the scoring rules presented in section 2.1. For the JF mechanism, the function $R(s_l, s_j)$ is 1 if $s_l = s_j$ and 0 otherwise.

A strategy σ is a Nash equilibrium strategy if and only if for all observable signals s_i, the agent a does not have the incentive to deviate from σ given that the rater also uses the reporting strategy σ:

$$E[\sigma, \sigma, s_i] \geq E[\sigma', \sigma, s_i] \qquad \text{for all} \quad s_i \in S, \sigma' \neq \sigma; \tag{1}$$

The MRZ mechanism generally has many Nash equilibrium strategies. Taking again the example in Figure 1, the reporting strategy $+$ (i.e. always reporting $+$) is also a Nash equilibrium: given that a_r reports $+$, it is best for a to also report $+$ since $R(+,+) = 0.95 > R(+,-) = 0.24$. Moreover, the expected payoff to a from the reporting strategy $+$ is greater than the expected payoff of the incentive-compatible strategy $(0.95 > 0.79)$.

The JF mechanism suffers from the same drawback. Reporting only $+$, reporting only $-$ or always reporting the opposite signal are all Nash equilibrium strategies that generate higher payoffs that the incentive-compatible strategy.

3.2 The influence of trusted reports

For all incentive compatible reputation mechanisms, the truthful reporting strategy σ^* is a strict Nash equilibrium. When the report submitted by the rater is always a trusted report, the expected payment received by a, given that she has observed the signal s_i and uses the reporting strategy σ is $E[\sigma, \sigma^*, s_i]$. As the rater's strategy is fixed, the only Nash equilibrium strategy of a is the truthful reporting strategy σ^*. Any other reporting strategy will generate a strictly lower payoff (Equation 1).

Since trusted reports are expensive, it is interesting to see if undesired Nash equilibrium points can be eliminated by using only a probabilistic comparison with a trusted report. Let q be the probability that the report of the rater is a trusted one. The expected payoff to a from the equilibrium strategy σ, given that she has observed the signal s_i is then:

$$E_q[\sigma, \sigma, s_i] = q \cdot E[\sigma, \sigma^*, s_i] + (1-q) \cdot E[\sigma, \sigma, s_i]$$

The strategy σ continues to be a Nash equilibrium strategy if and only if for all other reporting strategies σ', $E_q[\sigma', \sigma, s_i] < E_q[\sigma, \sigma, s_i]$, for all signals s_i.

Finding the minimum probability q such that the incentive-compatible reporting strategy remains the only Nash equilibrium strategy of the mechanism involves solving the following problem:

Problem 1. Find $q^* \in [0,1]$ such that for all q, $q^* \leq q \leq 1$, for all reporting strategies $\sigma \neq \sigma^*$, there is a signal s_i and a strategy $\sigma' \neq \sigma$ such that $E_q[\sigma, \sigma, s_i] < E_q[\sigma', \sigma, s_i]$.

In other words, Problem 1 implies finding the minimum probability with which the rater's report has to be true (i.e. trusted) such that for all reporting strategies σ, there is at least one profitable deviation (defined by the strategy σ') for at least one of the observable signals.

Problem 1 is very hard to solve for the general case, and the result is also very restrictive. A relaxation would be to eliminate only those equilibrium strategies that generate a higher payoff than the incentive compatible strategy. The practical justification for this relaxation is that rational agents always choose from a set of possible equilibrium strategies the one which generates the highest payoff. Therefore, given that truthful reporting yields the highest payoff, we argue that it is not necessary from a practical perspective to eliminate all other Nash equilibrium points.

Finding the minimum probability such that the incentive-compatible reporting strategy generates the highest payoff implies solving the following problem:

Problem 2. Find $q^* \in [0,1]$ such that for all q, $q^* \leq q \leq 1$, for all Nash equilibrium reporting strategies $\sigma \neq \sigma^*$, $E_q[\sigma, \sigma, s_i] < E_q[\sigma^*, \sigma^*, s_i]$, for all $s_i \in S$.

which can be reformulated as the following optimization problem:

Problem 3. *Minimize*: q
Under the constraint: $f(q) \geq 0$;
Where $f(q)$ is itself the optimization problem:
Maximize: $f(q) = E_q[\sigma, \sigma, s_i] - E[\sigma^*, \sigma^*, s_i]$
Under NE constr.: $E_q[\sigma, \sigma, s_k] \geq E_q[s_j, \sigma, s_k]$ for all $s_j, s_k \in S$.

The above optimization problem involves solving two nested optimizations: (1) finding the Nash equilibrium strategy that generates the highest payoff, and (2) finding the minimum value of q (i.e. q^*) for which the highest Nash equilibrium payoff corresponds to the incentive-compatible reporting strategy. Finding the highest Nash equilibrium payoff is a NP-hard problem [4]. The function $f(q)$, on the other hand, is increasing in q and therefore a binary search can be used to find the minimum value of q. Please note that the solutions to problem 3 also represent lower bounds for the solutions of problem 1.

The threshold value q^* for the probability that the report of the rater needs to be trusted is not necessarily the overall percentage of trusted reports needed by the reputation mechanism.

The structure of the MRZ mechanism allows to reuse trusted reports. The same rater (and report) can be used to assess the honesty of more than one feedback. In extremis, one could imagine that a trusted report is used to assess all other reports collected by the reputation mechanism. The actual percentage of trusted reports needed by the mechanism is hence very low and is equal to the value of q^* divided by the total number of reports which are rated against the same trusted report.

Rating more reports against the same trusted report poses however, some problems. The evaluation of the reports has to be done in the same time, after all reports have been submitted. The bigger the number of reports that are scored against the same trusted report, the smaller the overall percentage of trusted reports needed, but also the bigger the waiting time for the reputation side-payment. When agents discount future payoffs, the maximum waiting time becomes bounded (bigger waiting time will cancel the incentives to report the truth due to the discount factor).

Moreover, in a dynamic system (i.e. the type of the service might change in time), the trusted report could become outdated. Thus, a compromise needs to be reached between an increased cost due to higher percentage of trusted reports and a bigger risk of loosing the incentives to report the truth due to waiting times and outdated reports used for rating.

The JF mechanism, on the other hand requires a fresh rater for every submitted report (i.e. the report of the next agent is always used to rate the present feedback). Therefore, the threshold value q^* becomes the overall percentage of trusted reports needed by the reputation mechanism.

4 Numerical Analysis

In this section we numerically analyze the influence of trusted reports for a series of scenarios which are likely to appear in real reputation systems. For each such scenario we will present the numerical solution of the optimization problem 3, representing the threshold value (q^*) for the probability that the rater needs to provide a trusted report.

For the JF mechanism there is a closed form solution to the optimization problem 3:

$$q \geq q^* = \max\left(\frac{1 - Pr[-|-]}{Pr[-|-]}, \frac{1 - Pr[+|+]}{Pr[+|+]}\right);$$

As the probabilities $Pr[-|-]$ and $Pr[+|+]$ vary in the interval $[0.5, 1]$ (smaller values for $Pr[-|-]$ or $Pr[+|+]$ are not allowed by the assumptions of the JF mechanism), q^* takes values in the interval $[0, 1]$.

The threshold value (i.e. q^*) approaches 0 when both probabilities $Pr[+|+]$ and $Pr[-|-]$ approach 1. This scenario appears when the behavior of the provider in successive transactions is highly correlated. In such cases, there is little uncertainty about the signal observed by the rater, and therefore the incentive-compatible strategy of the JF mechanism yields payoffs which approach the maximum possible payoff.

For the MRZ mechanism we investigate settings with N possible types and N possible signals received by the agents. We considered that each signal corresponds to one type, and that there is uniform noise in the observation of signals. Let s_i be the signal corresponding to the type t_i. The probability distribution of signals conditioned on the type t_i is:

$$f(s_j|t_i) = \begin{cases} 1 - \delta & \text{if} \quad s_j = s_i \\ \dfrac{\delta}{N-1} & \text{if} \quad s_j \neq s_i \end{cases}$$

where δ is the “level” of noise, typically, $\delta = 10\%$.

For $N = 2$, Figure 2 plots the threshold value q^* for all possible beliefs of the agents. Having only 2 types for the service (i.e. *good* and *bad*) the probability of the type being good (the value on the horizontal axis) completely describes a belief.

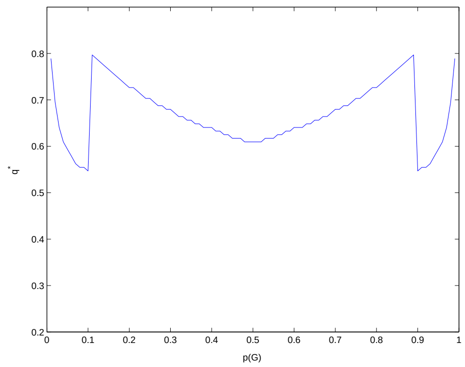

Fig. 2. Threshold value q^* for the MRZ mechanism, when $N = 2$ and $\delta = 10\%$.

From Figure 2 one can see that the report of the rater has to be trusted with a probability between 60 and 80 percent. The gaps at both ends of the interval are explained by the "activation" of previously inefficient reporting strategies. When the probability of the type G is close to one, the constant reporting strategy $-$ is very inefficient. The $Pr[-|-]$ has a very small value, which plugged into the formula of any of the scoring rules presented in Section 2.1 yields a very small payoff. Therefore, the strategy of always reporting $-$, though still a Nash equilibrium, is not "activated" unless the prior probability of the B type crosses a certain threshold. This behavior provides an important insight into the influence of trusted reports: Trusted reports have a big impact on eliminating individual lying strategies (the exponential decrease at the end of the interval when only one of the constant reporting strategies is "active"). However, they cannot simultaneously eliminate all of them.

For $N = 3$ the space of possible beliefs is two dimensional: two probabilities entirely characterizes the prior distribution over the three types. Figure 3 presents three slices through the 3-dimensional graph for three different probabilities of the type t_1: $p(t_1) = 0.1$, $p(t_1) = 0.3$ and $p(t_1) = 0.5$. The threshold value q^* varies between 0.5 and 0.8. Higher values characterize more focused beliefs (i.e. beliefs that are highly focused around one type) while lower values characterize more ambiguous beliefs (i.e. beliefs for which the probability distribution over types is more flat).

For higher number of types, the graphical representation of the space of beliefs becomes impossible. Moreover, solving the optimization problem for an increasing number of types (and therefore signals) becomes exponentially more difficult. Instead of determining the threshold values for all possible beliefs, we have concentrated on a smaller set of beliefs with increased practical importance. For every $N \in \{4, 5, 6\}$ we took all of the beliefs according to which the prior probability of types is normally distributed around one of the types. The table in Figure 4 summarizes the threshold values q^* for the probability that the report of the rater needs to be trusted.

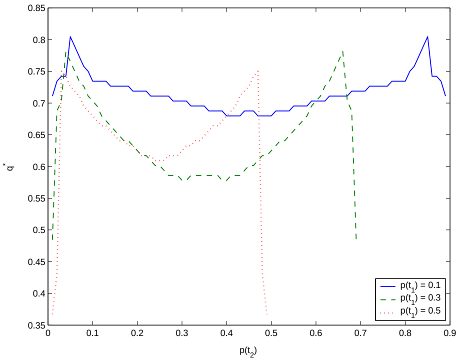

Fig. 3. Threshold value q^* for the MRZ mechanism, when $N = 3$ types, and $\delta = 10\%$.

	Belief normally distributed around:					
	t_1	t_2	t_3	t_4	t_5	t_6
$N = 4$	0.664	0.539	0.539	0.664	-	-
$N = 5$	0.675	0.734	0.734	0.734	0.750	-
$N = 6$	0.679	0.664	0.664	0.656	0.812	0.398

Fig. 4. Threshold value q^* for the MRZ mechanism, when $N \in \{4, 5, 6\}$ and $\delta = 10\%$.

All values have been produced using the global optimization solver BARON[6] and the mathematical modelling language AIMMS[7]. For lack of computational time, we couldn't investigate the threshold values for settings with more than 6 types.

5 Related work

The formal study of reputation mechanisms was started by the three seminal papers by Kreps, Wilson, Milgrom and Roberts ([13, 14, 12]) who proved that cooperative equilibria can exist in finitely repeated games due to the reputation effect. Ever since, reputation mechanisms have received an increasing amount of interest, both from a practical and theoretical point of view.

Examples of computational trust mechanisms based on reputation are numerous, ranging from mechanisms based on direct interactions (in [2] agents learn to trust each other by keeping track of past interactions) to complex social networks [17] in which agents ask and give recommendations to their peers. Centralized implementations as well as completely decentralized [1] have been investigated.

One major challenge associated with designing reputation mechanisms is to ensure that truthful reports are gathered about the actual outcome of the transaction. Besides the two solutions described in this paper, there have been a couple of other incentive compatible mechanisms.

[6] http://archimedes.scs.uiuc.edu/baron/baron.html

[7] http://www.aimms.com

An interesting approach is that of Braynov and Sandholm in [3] and Dellarocas in [6]. Instead of making it rational for the reporters to provide honest feedback, the authors incentivize the service providers to truthfully declare their trustworthiness (or reputation).

[3] considers exchanges of goods for money and proves that a market in which agents are trusted to the degree they deserve to be trusted is equally efficient as a market with complete trustworthiness. By scaling the amount of the traded product, the authors prove that it is possible to make it rational for sellers to truthfully declare their trustworthiness. Truthful declaration of one's trustworthiness eliminates the need of reputation mechanisms and significantly reduces the cost of trust management.

For e-Bay-like auctions, the Goodwill Hunting mechanism [6] provides a way to make the sellers indifferent between lying or truthfully declaring the quality of the good offered for sale. Momentary gains or losses obtained from misrepresenting the good's quality are later compensated by the mechanism which has the power to modify the announcement of the seller.

Finally, Jurca and Faltings [11] take a different approach and achieve in equilibrium truthful reporting by comparing the two reports coming from the buyer and the seller involved in the same transaction.

This paper also relates to the vast literature concerned with computing Nash equilibrium strategies. [4] gives a nice overview of the available results in this field and proves some complexity results which are of direct relevance to the present paper.

Last, but not least, this work relates to the ongoing efforts of the networking community to design routing algorithms that have a unique Nash equilibrium point with the desired properties.

6 Conclusion

In this paper we analyze the influence of trusted reports on the set of Nash equilibria of two well-established incentive-compatible reputation mechanisms. We emphasize the problem such mechanisms have with non-incentive compatible Nash equilibrium strategies, and investigate how such undesired equilibrium points can be eliminated. By using trusted reports to rate the honesty of feedback submitted by rational agents we show that it is possible to have a mechanism in which the incentive compatible strategy is the only (or the most likely) strategy to be followed.

A numerical analysis provides values for the minimum probability that the report of the rater needs to be trusted. For the JF mechanism, this threshold value can be smaller; it corresponds, however, to the overall percentage of trusted reports needed by the mechanism. The MRZ mechanism requires a higher threshold values, but on the other hand allows the reuse of trusted reports.

Besides the numerical analysis of the two reputation mechanisms we also provide a general methodology for eliminating undesired equilibrium points from incentive compatible reputation mechanisms.

References

1. K. Aberer and Z. Despotovic. Managing Trust in a Peer-2-Peer Information System. In *Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM)*, 2001.
2. 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.
3. S. Braynov and T. Sandholm. Incentive Compatible Mechanism for Trust Revelation. In *Proceedings of the AAMAS*, Bologna, Italy, 2002.
4. V. Conitzer and T. Sandholm. Complexity Results about Nash Equilibria. In *Proceedings of the IJCAI*, Acapulco, Mexico, 2003.
5. R. Conte and M. Paolucci. *Reputation in Artificial Societies. Social Beliefs for Social Order.* Kluwer, Boston (MA), 2002.
6. 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.
7. C. Dellarocas. The Digitization of Word-of-Mouth: Promise and Challenges of Online Feedback Mechanisms. MIT Sloan Working Paper No. 4296-03., 2003.
8. C. Dellarocas. Sanctioning Reputation Mechanisms in Online Trading Environments with Moral Hazard. MIT Sloan Working Paper #4297-03, 2004.
9. D. Houser and J. Wooders. Reputation in Internet Auctions: Theory and Evidence from eBay. University of Arizona Working Paper #00-01, 2001.
10. R. Jurca and B. Faltings. An Incentive-Compatible Reputation Mechanism. In *Proceedings of the IEEE Conference on E-Commerce*, Newport Beach, CA, USA, 2003.
11. R. Jurca and B. Faltings. “CONFESS”. An Incentive Compatible Reputation Mechanism for the Online Hotel Booking Industry. In *Proceedings of the IEEE Conference on E-Commerce*, San Diego, CA, USA, 2004.
12. 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.
13. D. M. Kreps and R. Wilson. Reputation and Imperfect Information. *Journal of Economic Theory*, 27:253–279, 1982.
14. P. Milgrom and J. Roberts. Predation, Reputation and Entry Deterrence. *J. Econ. Theory*, 27:280–312, 1982.
15. N. Miller, P. Resnick, and R. Zeckhauser. Eliciting Informative Feedback: The Peer-Prediction Method. Forthcoming in Management Science, 2005.
16. P. Resnick and R. Zeckhauser. Trust Among Strangers in Electronic Transactions: Empirical Analysis of eBay’s Reputation System. In M. Baye, editor, *The Economics of the Internet and E-Commerce*, volume 11 of Advances in Applied Microeconomics. Elsevier Science, Amsterdam, 2002.
17. B. Yu and M. Singh. Detecting Deception in Reputation Management. In *Proceedings of the AAMAS*, Melbourne, Australia, 2003.