"CONFESS". An Incentive Compatible Reputation Mechanism for the Online Hotel Booking Industry.

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

Abstract

Reputation mechanisms provide a promising alternative to the traditional security methods for preventing malicious behavior in online transactions. However, obtaining correct reputation information is not trivial. In the absence of objective authorities (or trusted third parties) which can oversee every transaction, mechanism designers have to ensure that it is rational for the participating parties to report the truth. In this paper we describe a complete reputation mechanism for the online hotel booking industry that is efficient (i.e. the equilibrium behavior is cooperative) and incentive compatible. Our mechanism discovers the true outcome of an interaction by analyzing the two reports coming from the agents involved in the interaction. Based on side payments, such a mechanism makes it profitable for long-run agents to commit to always report the truth.

1. Introduction

Trust and reputation mechanisms provide a promising alternative to the traditional methods used to prevent misbehavior in business transactions. In an open, heterogeneous, and often mobile environment, old security mechanisms involving strict laws and enforcing authorities are very hard to deploy. Moreover, when business is conducted through software programs acting on behalf of humans (so called agents), the decreased level of physical interaction leaves the system much more susceptible to fraud and deception. Numerous examples of online commerce fraud attest to the fact that the old idea of control should be replaced by more robust and flexible mechanisms that can ensure the necessary level of trust essential to the functioning of every market.

Reputation mechanisms 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.

One major challenge associated with designing reputation mechanisms is to ensure that truthful information is gathered about the actual outcome of the transaction. In the absence of independent verification means, a reputation mechanism has to rely on the information provided by the parties involved, information distorted by the strategic interests of the reporters.

In most scenarios, reporting the truth is not the rational thing to do, and therefore should not be expected from autonomous, utility maximizing agents. Let us consider a typical consumer-provider setting in which the consumer is completely trustworthy, however, the trustworthiness of the provider is questionable. A true positive report could create inconveniences for the consumer: the increased reputation could attract other consumers and decrease the future availability of the provider for the reporting agent. Moreover, in a competitive environment, falsely submitting a negative report slightly increases the reporter's reputation with respect to the others.

Incentive compatibility (i.e. making it rational for consumers to report the truth) can be achieved by side payments that reward a reputation report of a consumer proportionally to the correlation with future, unknown reports (assumed to be true), about the same provider. For specific environments, [12] and [7] describe such schemes that make truth revelation a Nash equilibrium.

In this paper we present an incentive-compatible reputation mechanism for a more general setting in which both parties involved in the transaction are assumed to behave rationally. While for the provider it is beneficial to have a good reputation for providing quality goods or services, we show that in certain circumstances it is beneficial for consumers to have a good reputation for always reporting the truth. The mechanism is based on the observation that providers are less likely to cheat on consumers that are probably going to report that defection (i.e. have a good reputation for reporting the truth) as the resulting negative report will attract future losses that outweigh the momentary gain obtained from cheating.

The mechanism uses a *semantically well defined* notion of reputation, and obtains truthful feedback by analyzing the two reports coming from the consumer and provider involved. The novelty of this mechanism consists in the fact that we allow the provider to confess defection, before asking the consumer to provide feedback. In a different paper, [8], we prove that this interaction protocol allows consumers to build a reputation for correctly reporting the behavior of the provider.

The mechanism we are going to present (named "CONFESS") is tailored to the realistic environment of online hotel booking industry. Section 2 presents the scenario in which we describe our mechanism, Section 3 defines the notion of reputation, and Section 4 presents the full reputation mechanism. Finally, we compare our mechanism with related work, and conclude.

2. The Scenario

We consider the example of one hotel having N rooms that offer exactly the same accommodation conditions. The quality of the hotel is judged by taking into consideration a number of criteria, e.g. the level of noise, cleanness, available facilities, the professionalism of the staff, etc. We make the simplifying assumptions that the values of all these attributes can be combined into one measure of the quality of the service offered by the hotel.

Let us use in the rest of this paper a normalized value for the quality of service of the hotel, such that a quality of 1, denotes the best possible service offered by any hotel. Similarly, a quality of 0 denotes the worst possible service. It is common knowledge in the environment that any customer is willing to pay w dollars for a night spent in a hotel offering the best possible service (i.e. quality of service 1). We define the real quality of service, α, of a particular hotel, such that a customer who knows the service of that hotel is willing to pay αw dollars for one room. We also assume that all customers, given enough information, agree on the same number for the quality of service of one hotel.

The management of each hotel decides upfront the investment I it is going to make for building the hotel. This investment determines the available space, the quality of interior decorations, the training of the personnel, the available facilities etc. We assume that investment I uniquely determines α, the maximum attainable quality of service offered by that hotel. However, in order to actually provide the quality α, the hotel has to spend w_r dollars every night for every occupied room. w_r should be regarded as running costs and includes the room cleaning, room service, appropriate curtesy and support to the client, maintaining working elevators and phones, etc. The hotel, also decides upfront the quality level β it is going to advertise for its rooms.

The hotel can decide every night for every client whether or not to spend w_r. If w_r is spent (i.e. hotel cooperates), the client will experience the maximum quality level α; if the hotel doesn't spend w_r (i.e. hotel cheats), the client will experience a quality level much lower than α. The client is happy when she receives the promised quality level (i.e. what she paid for) but feels cheated whenever she receives anything less. A happy client gets a payoff of $\rho_C \alpha w$ for every interaction (i.e. night spent in the hotel) while a dissatisfied client gets a payoff of $-\alpha w$, i.e. she feels she threw away the money paid for the room.

After every night, each client submits a binary report about the hotel: a positive report (1) if she was happy with the service of the hotel, or a negative report (0) if she felt cheated. All the reports are aggregated by a Reputation Mechanism (RM) into one measure $R \in [0, 1]$, of the reputation of the hotel. The RM will be presented in detail in Sections 3 and 4.

We assume that the reputation of the hotel directly affects its occupancy. If R_t is the reputation of the hotel at night t, the hotel will occupy:

$$O_t = N[(1 - a) \cdot R_t + a]$$

of its rooms in that night. $a \in [0, 1]$ is the percentage of rooms which are occupied independently of the reputation of the hotel. a reflects the customers who don't have a choice, or who do not check the reputation of the hotel.

The dependence of the occupation rate on the reputation of the hotel is just a way of modeling the influence of present reputation on future gains. As later shown in Section 3, this assumption is essential for correctly defining the notion of reputation. Other forms of dependence between the revenue of the hotel and its reputation can be implemented by modifying the definition of reputation accordingly. The "CONFESS" mechanism described in this paper is independent of this assumption, and will work for other definitions of reputation as well.

The per night revenue of the hotel is:

$$g(R_t, r_t) = N[(1 - a)R_t + a]\beta w - r_t w_r; \qquad (1)$$

where:

- r_t is the number of rooms for which the hotel invested the running costs w_r. The value of r_t lies between 0 and $N[(1-a)R_t + a]$, the number of occupied rooms;
- R_t is the reputation of the hotel at night t.

Assuming that the hotel's payoff is discounted with the daily discount factor δ, the average payoff of the hotel is:

$$V = (1-\delta)\sum_{t=0}^{\infty} \delta^t g(R_t, r_t); \quad (2)$$

From Equations (1) and (2) we see that a hotel might cheat a naive customer by (a) declaring a higher initial quality of service and (b) refusing to spend the daily running cost w_r, necessary to fulfill the customer's expectations. The role of a reputation mechanism is to ensure that the hotel will have the incentive to behave cooperatively: i.e. declare the true quality of its rooms and spend the necessary amount for the running costs. This can be done by making sure that misbehavior in the present will attract a penalty in the future revenues due to a bad reputation. If the future penalty outweighs the short term gain obtained from cheating, a rational hotel will never cheat on its customers.

3. Semantics of Reputation

The Reputation Mechanism presented here is an extension of the one described by Dellarocas in [5]. The reputation of a hotel is stored as a set S of cardinality M of binary reports (i.e. 0 or 1) where M is one parameter under the control of the mechanism designer. Every new submitted report replaces one randomly selected report from S. The randomization of the reputation updating process eliminates any advantage a hotel might obtain from knowing the exact sequence of past reputation reports. The reputation information R made public to the potential customers is the fraction of the positive reputation reports from S, i.e.:

$$R = \frac{\textit{number of positive reports in } S}{M};$$

The reputation of the hotel is updated after every night in the following way: From S, N reports are taken at random and deleted. The deleted reports are replaced with the ones submitted by the customers from that night. A negative report is also submitted for every empty room.

Let $f : [0,1] \times \{0,1\} \rightarrow [0,1]$ be the reputation update function corresponding to the rule enounced above, such that $R_{t+1} = f(R_t, r_t)$, where R_t is the reputation of the hotel at time t and $r_t \in [0, N((1-a)R_t + a)]$ is the number of rooms for which the hotel invested w_r, (equal to the number of positive reputation reports submitted about the hotel) at night t. The expected value of the updated reputation is therefore:

$$R_{t+1} = E[f(R_t, r_t)] = R_t\left(1 - \frac{N}{M}\right) + \frac{r_t}{M}; \quad (3)$$

The goal of the RM is to make the hotel (a) declare its true quality of service and (b) spend the running costs w_r every night.

Theorem 1 *When the hotel truthfully declares its quality of service, all customers submit truthful feedback, and:*

$$\frac{\delta(1-a)N}{M(1-\delta)+\delta N} > \frac{w_r}{\alpha w};$$

the hotel maximizes its revenues by cooperating (i.e. spending the running costs w_r) every night, for every client.

PROOF. A rational hotel chooses the actions (i.e. the sequence of numbers r_t) which maximize its lifetime revenue:

$$V^{max} = \max_{(r_t)} V = \max_{(r_t)} (1-\delta)\sum_{t=0}^{\infty} \delta^t g(R_t, r_t); \quad (4)$$

where R_0 is a constant. Equation (4) is in Bellman form, and therefore, a control sequence (r_t^*) is optimal if it is *unimprovable*, i.e. there is no profitable one stage deviation from (r_t^*).

Let (r_t^*) be the cooperative action sequence of the hotel, i.e. the hotel invests w_r every night for every customer. Therefore, $r_t^* = N[(1-a)R_t^* + a]$ for all t, and (R_t^*) is the sequence of values of the hotel's reputation generated by Equation (3). Let us also consider the action sequence (r_t) being the smallest one stage deviation from (r_t^*): $r_t = r_t^*$ for all $t \neq \tau$, and $r_\tau = r_\tau^* - 1$. (R_t) is the corresponding sequence of values for the reputation.

$$V^* - V = (1-\delta)\sum_{t=0}^{\infty} \delta^t (g(R_t^*, r_t^*) - g(R_t, r_t)); \quad (5)$$

By replacing the definitions of (r_t^*) and (r_t), by using Equations (1), (3) and the assumptions of the theorem, we obtain $V^* - V > 0$ which means that (r_t^*) is unimprovable, and therefore optimal. ■

The conditions under which the hotel has the incentive to truthfully declare its quality of service are described in the following theorem:

Theorem 2 *When customers submit truthful feedback, and:*

$$\alpha > \frac{w_r}{w} + \frac{M(1-\delta) + a\delta N}{M(1-\delta)+\delta N};$$

it is in the best interest of a hotel to declare its true quality of service.

PROOF. Starting from the intuitive assumption that it is not rational for any hotel to declare a quality of service smaller than the real one, in the rest of this proof we will show that exaggerating the quality of service will make the hotel lose money in the long run.

Let us suppose that the hotel will declare a quality of service β greater than the real one, α. No mater how hard it tries, the hotel will never be able to fulfil its promises: the hotel promises quality level β, but can provide at most quality level $\alpha < \beta$. The clients will always be dissatisfied and therefore submit negative reputation reports. Since the hotel will anyway receive only negative reputation reports, it will not invest anything in the running costs. Therefore, its overall payoff will be:

$$V = (1-\delta)\sum_{t=0}^{\infty}\delta^t g(R_t, 0); \quad R_t = R_0(1-N/M)^t;$$

using Equation (3), where R_0 is the given reputation of the hotel at time 0.

Let us compare this payoff V, with the payoff V^*, of a hotel who declares its true quality of service, and always spends the running costs for every customer. We know from Theorem 1 that V^* is the maximum payoff the hotel can get when truthfully declaring its quality of service. As in this case a hotel will always receive a positive reputation report,

$$V^* = (1-\delta)\sum_{t=0}^{\infty}\delta^t g(R_t^*, r_t^*); \quad R_t^* = 1+\left(1-\frac{aN}{M}\right)^t (R_0-1);$$

By making the necessary replacements and using the assumptions in the theorem, we obtain $V^* - V > 0$. ■

The properties given above are based on the assumption that customers submit true reputation reports. This is not however a realistic assumption. For example, we might consider that a hotel with a better reputation is less likely to have free rooms. A rational customer will therefore be slightly biased towards providing negative feedback, which will increase her future chances of booking a room.

The next section presents an integrated reputation mechanism that is incentive compatible.

4. Incentive Compatible Reputation Mechanism

The reputation mechanism we are presenting here is adapted from [8] in order to meet the requirements of the environment described in this paper. The mechanism is based on the key assumption that hotel customers can be modeled as long-run players (e.g. business travelers usually return to the same hotel if they are happy with the service provided), and it functions according to the intuition that long run customers can benefit from building a reputation as honest reporters. The hotel will fear cheating on a customer who has a reputation for always reporting the truth because the resulting negative report will affect the hotel's future revenues. The customer will therefore obtain the promised service which pays for the effort invested in building the reputation.

The novelty of this mechanism is that it also requires the hotel to *confess* its behavior relative to a particular client. The hotel can submit a positive report (1) if it claims having cooperated, or a negative report (0) to admit having defected. By correlating the report of the hotel with that of the client about the same interaction, three cases are possible:

1. The hotel admits having cheated. For a hotel, falsely acknowledging defection implies a double loss (i.e. the future loss due to a negative reputation report, and the momentary loss coming from not taking the opportunity of defecting) and therefore no rational hotel will report 0 without actually defecting. Regardless of the client's report, we can conclude in this case that the hotel has indeed cheated.

2. Both parties report 1. The interaction was most likely cooperative in this case, and therefore a positive report can be recorded for the hotel.

3. The hotel claims cooperative behavior while the client reports a negative report. In this case, we know that one of the agents is surely lying. Since untruthful reporting is what we seek to avoid, both the hotel and the client will be punished in this case: a negative report is being recorded for the hotel, and both parties are fined for lying.

The mechanism requires the existence of a central entity (the center) capable of charging fees from both the hotel and the customer. A booking site like Expedia[1] or Hotels.com[2] can easily play this role.

The interaction protocol goes as follows. Every night, the center advertises the N rooms of the hotel, each at a price of αw dollars a night. The center charges the hotel a listing fee ε_H for every room booked by a client. Each client pays αw dollars to the hotel (through the center) and also pays the fee ε_C to the center. We assume that the hotel can decide every night, for every client whether or not to spend the running costs w_r necessary for delivering the promised quality of service. Moreover, we assume that the clients have no way of a priory knowing what the hotel decides.

Next morning, the center starts collecting reputation information about the behavior of the hotel. For every client from the previous night, the center runs the following protocol:

1. The hotel is first required to submit a report. If the hotel admits having cheated, a negative report is registered by the RM, and the fees ε_H and ε_C are returned to the rightful owners.

2. If however the hotel pretends to have cooperated, the client is asked to provide a report.

[1] www.expedia.com
[2] www.hotels.com

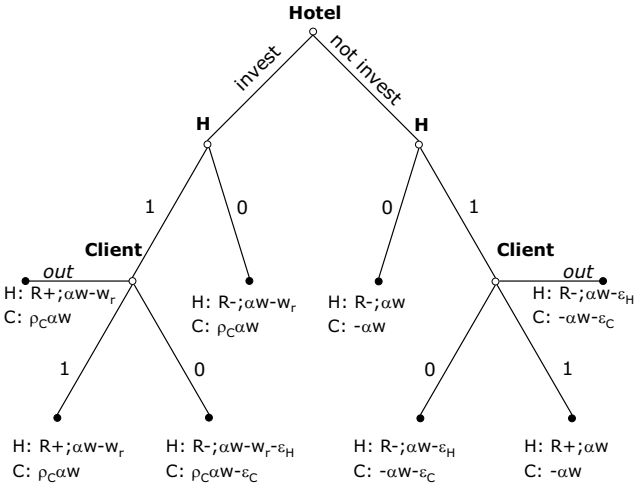

Figure 1. The Interaction Protocol.

3. If the client submits a positive report, the listing fees ε_H and ε_C are returned, and the RM records a positive report for the hotel.

4. If the client submits a negative report, the listing fees ε_H and ε_C are confiscated, however, the RM believes the client and registers a negative report for the hotel.

When the reports from all the clients are available, the RM updates the reputation of the hotel as presented in Section 3.

After every interaction, the client can decide never to come back to that hotel (i.e. take action out). In this case, we assume that she can book rooms at a similar hotel, which is entirely trustworthy, however more expensive: one room offering the quality level α costs $\alpha w(1+\theta)$, with $\theta > 0$.

Figure 1 presents a schematic view of the interaction protocol between the hotel and one client. The leaves indicate the per-interaction payoffs received by the hotel and the buyer respectively. The hotel's payoff is a tuple mentioning the reputation report ($R+$ or $R-$) registered by the RM, and the monetary gain for that interaction. Note that the repeated interaction between the hotel and one client can be isolated from the other interactions between the hotel and other clients by attributing a value to the reputation report submitted after every interaction. The value of the reputation report concentrates all influences that report has on the future revenue of the hotel. We can therefore analyze the interaction between the hotel and each client separately.

In [8], Jurca and Faltings present a game theoretic analysis of the above described protocol. When agents have game-theoretic perfect information (i.e. both the hotel and the client are rational, and this fact is common knowledge), there is an upper bound on the percentage of false reports registered by the RM. This upper bound is given by:

$$\overline{p} = \frac{(1-\delta_C)\varepsilon_C + \delta_C \alpha w \theta}{\delta_C \alpha w (1+\rho_C)}; \tag{6}$$

where δ_C is the discount factor of the client.

The performance of the mechanism is greatly improved if we introduce some small amount of uncertainty in the game. We assume that the hotel is not perfectly informed about the type of the client, i.e. the hotel believes with some prior probability μ_0^* that the client has a different payoff structure which makes her prefer to always report the truth about the behavior of the hotel.

In such a setting, [8] proves that it is possible for the clients to build a reputation for always reporting the true behavior of the hotel. The idea behind this result is the following. When a short sighted client is cheated by a hotel which afterwards claims having cooperated, (i.e. hotel reports 1) the client will also submit a positive report (i.e. client reports 1) in order to avoid the fine ε_C for lying. A log-run client, might however use the following reasoning: By reporting 0 when the hotel cheated but claimed having cooperated, the client looses the fine ε_C, but also transmits a message to the hotel. This apparently irrational behavior of the client makes the hotel believe that any future defection will likely to be exposed by the client. In other words, the client invests in a reputation for always reporting the truth.

The fact that the client builds a reputation for truthful reporting affects the behavior of the hotel. When dealing with a reputable client, it is not rational for the hotel to cheat: defection will most likely be exposed, and from the definition of reputation, the resulting negative reputation report offsets the momentary gain obtained from cheating. A rational hotel will therefore treat correctly all reputable customers. This in turn generates a higher future payoff for the reputable customers which pays for the investment in the reputation.

The above mentioned reputation effect imposes an upper bound on the number of rounds in which the hotel will cheat on a client who always reports the truth. In [8], Theorem 1 gives a closed form solution for this upper bound k:

$$k = \left\lceil \frac{\ln(\mu_0^*)}{\ln(\overline{\pi})} \right\rceil; \tag{7}$$

where μ_0^* is the prior probability of the hotel's belief that a customer will always report the truth (i.e. the customer is committed to always report the truth) and:

$$\overline{\pi} = \frac{(1-\delta_H)w_r + \delta_H \Phi}{(1-\delta_H)\left[w_r + \epsilon + \varepsilon_H\right] + \delta_H \Phi};$$

where:

- δ_H is the discount factor of the hotel for the interaction with a particular client. Note that there is a difference between δ and δ_H. If for example a client will come to the hotel once a year (i.e. every 365 days), $\delta_H = \delta^{365}$;

- ε_H is the lying fine for the hotel which can be tuned by the mechanism designer;
- $\epsilon = V(t, r_t = r^*) - V(t, r_t = r^* - 1)$ is the loss of the hotel as a consequence of receiving one negative reputation report instead of a positive one; By replacing (2) and (3) we obtain:

$$\epsilon = \frac{\delta(1-a)N\alpha w}{M(1-\delta)+\delta N} - w_r;$$

- Φ describes the maximum difference between any two payoffs the hotel can obtain when interacting with a rational customer.

$$\Phi = w_r \frac{(1-\delta_C)\varepsilon_C + \delta_C \alpha w \theta}{\delta_C \alpha w (1+\rho_C)};$$

The existence of the upper bound k, further reduces the probability with which the RM will accept a false reputation report. This new bound on the probability is given by:

$$\overline{p'} = \frac{(1-\delta_C)\varepsilon_C + (\delta_C - \delta_C^k)(\alpha w + \varepsilon_C + \alpha w \rho_C)}{\delta_C \alpha w (1+\rho_C)}; \quad (8)$$

From Equations (6) and (8), the upper bound on the percentage of false reputation reports accepted by the RM is given by:

$$p < \min(\overline{p'}, \overline{p})$$

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

$$\overline{p'} = \frac{(1-\delta_C)\varepsilon_C}{\delta_C \alpha w (1+\rho_C)};$$

and as ε_C 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.

Two properties of the mechanism are straight forward to prove:

Property 1 *The mechanism is bounded socially efficient.*

SKETCH OF PROOF. Every time the hotel does not cooperate, there is a social loss equal to $\alpha w(1+\rho_C) - w_r$. Because k limits the number of interactions in which the hotel does not cooperate, the social loss is bounded above by $k \cdot [\alpha w(1+\rho_C) - w_r]$. ■

Property 2 *The mechanism is weakly budget balanced*

SKETCH OF 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 $\varepsilon_C + \varepsilon_H$. By introducing supplementary service fees, the mechanism can be easily transformed into one that yields profit to the center. ■

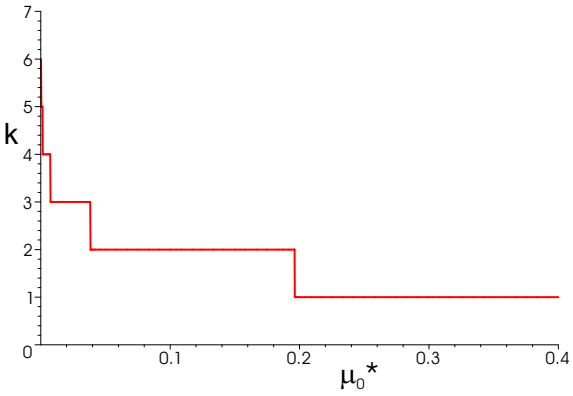

Figure 2. The upper bound k depending on the prior belief μ_0^*.

4.1. Numerical Example

When:

- $N = 20$ rooms
- $w = 200$ dollars for the perfect hotel room
- $\alpha = 0.7$ the quality of the hotel
- $a = 0.1$ the percentage of rooms which are occupied independent of the hotel's reputation
- $\delta = 0.9999$ the daily discount factor of the hotel. We assume that the client returns once a year with probability $\delta_C = 0.7$. $\delta_H = \delta^{365} \simeq 0.96$ is the yearly discount factor of the hotel
- $w_r = 10$ dollars running costs per night, per room
- $M = 500$ reputation reports kept in the set S
- $\varepsilon_H = 20$ and $\varepsilon_C = 1$ are the lying fines for the hotel and the client respectively.

Figure 2 plots the value of the upper bound k for different values of the prior belief, μ_0^*. Figure 3 plots the values of the probability bounds $\overline{p}$ and $\overline{p'}$ for different values of the prior belief, μ_0^*.

The values chosen for μ_0^* correspond to typical rates of cooperative behavior encountered for humans. As it can be seen, for $\mu = 0.2$, k equals to 1, and therefore it is rational for the hotel to cheat at most one time on any client. Because of this, the maximum probability with which the RM will register a false reputation report is 0.2%. Moreover, this probability can be further decrease, by decreasing ε_C. As ε_C approaches 0, the RM will never register a false report (in equilibrium).

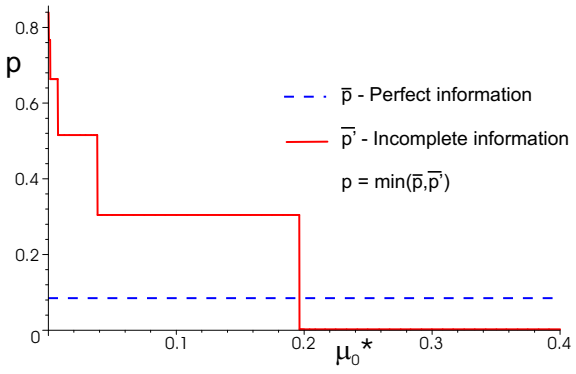

Figure 3. The maximum probability of registering a false report depending on the prior belief μ_0^*

4.2. Open Issues

As a client's reputation report does not affect the future terms of business between the hotel and that client (the price of the room remains the same), the only condition imposed on the client's fee ε_C is that it be strictly positive. There might be however other factors that encourage the customer to provide negative feedback. As mentioned in Section 3, we could consider that a present negative report increases the customer's chances of finding an available room in the future. ε_C could be used by the mechanism designer to counterbalance this bias. An exact form for ε_C is application dependent and requires a complete model of the influence of present reputation reports on future interactions.

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 hotel stores the reputation of the clients) we could have the same hotel and client interact through multiple centers (booking sites) without having to relay on one single centralized institution.

One direction of future research is to study the robustness of the mechanism to mistakes or imperfect monitoring of the hotel's actions. In the present form of the mechanism, a hotel's defection by mistake in a situation in which it was not rational for the hotel to defect will be interpreted by the clients as evidence of irrational behavior, and will invalidate the equilibrium results presented in the beginning of this section. Moreover, a client who falsely reported the reputation of the hotel once, will never again be able to build a credible reputation as a honest reporter. This problem become even more serious if we assume that the client does not perfectly perceive the action of the hotel.

Last but not least, the problem of collusion needs to be addressed. A hotel might create and interact with bogus clients in order to obtain an undeserved good reputation. Because the life of a negative report is limited (every negative report will at some point be deleted from the reputation set S) it is possible to imagine a scenario in which the hotel cheats on real customers, and than creates as many fictitious interactions as necessary in order to erase the negative reports. Ways in which this problem can be solved include:

- disregarding or underemphasizing the repeating reports coming from the same clients;
- charging fees for online identities;
- charging participation fees for every interaction.

5. Related Work

Theoretic research on reputation mechanisms started with the three seminal papers of Kreps, Milgom, Wilson and Roberts [9, 10, 11] who introduced the *reputation effect*, i.e. preference of agents to develop a reputation for a certain "type". Building a reputation however, involves some costs which have to be outrun by the future payoffs obtained when the reputation becomes credible. As a consequence, the reputation effect exists only in a certain class of games, with players meeting certain criteria.

Fudenberg and Levine [6] study the class of all repeated games in which a long-run player faces a sequence of single-shot opponents who can observe all previous games. Based on the reputation effect, the authors derive a lower bound on the payoff received by the long-run player in any Nash equilibrium of the repeated game. This result holds for both finitely and infinitely repeated games, and it is robust against further perturbations of the information structure.

Schmidt [13] provides a generalization of the above result for the two long-run player case in a special class of games called of "conflicting interests", when one of the players is sufficiently more patient than the opponent. The author derives an upper limit on the number of rounds player two will not play a best response to player one's commitment type, which in turn generates a lower bound on player one's equilibrium payoff. The same reasoning is used in [8] to prove the properties of the incentive-compatible reputation mechanism.

Computational trust mechanisms based on reputation are presented in [1, 2, 14]. Both direct (obtained from direct experience) and indirect (reported by other peers) reputation information is used, however, these mechanisms do not provide any rational incentives for the agents to participate.

Moreover, there is little protection against untruthful reporting, and no guarantee that the mechanism cannot be manipulated by a malicious provider in order to obtain higher payoffs.

Dellarocas [5] presents an efficient binary reputation mechanism that encourages a cooperative equilibrium in an environment of purely rational buyers and sellers. The mechanism is centralized, it works for single-value transactions, however, the buyers do not have any incentives to provide feedback. The same author addresses the problem of incentive-compatibility in [4].

A significant contribution towards eliciting honest reporting behavior is made in [12]. The authors propose scoring rules as payment functions which induce rational honest reporting. The scoring rules however, cannot be implemented without accurately knowing the parameters of the agents' behavior model, which can be a problem in real-world systems. Moreover, this mechanism can be used only when agents have typed behavior. Using the same principle, [7] overcomes the need to know the parameters of the agents' behavior model at the expense of further reducing the acceptable provider behavior types.

An interesting alternative is proposed in [3]. For auctions which are not completely enforceable, Braynov et al. describe a mechanism based on discriminatory bidding rules that separate trustworthy from untrustworthy bidders. This approach eliminates the need for reputation management, however, it is applicable only to some particular environments.

6. Conclusions and Future Work

In this paper we have presented an integrated reputation mechanism, adapted to the special conditions of the online hotel booking industry, that is both efficient (a cooperative solution is reached in equilibrium) and incentive-compatible. The mechanism is based on the more general assumption of agent rationality, and does not require the existence of a trusted third party to oversee every transaction. It is therefore easily adaptable to many other online commerce settings.

As presented above, "CONFESS" is a centralized mechanism. However, it is easily implementable in a distributed manner if we make the observation that the action of the center does not depend on any past or future information. A system can be build such that each interaction is directed by a different center that broadcasts the result of the transaction.

As future work, we plan to study the behavior of the above presented mechanism in the presence of mistakes (coming both from the hotel and from the customers) and irrationally malicious agents. We will also address the problem of collusion between the hotel and clients.

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] S. Braynov and T. Sandholm. Auctions with Untrustworthy Bidders. In *Proceedings of the IEEE Conference on E-Commerce*, Newport Beach, CA, USA, 2003.

[4] 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.

[5] C. Dellarocas. Efficiency and Robustness of Binary Feedback Mechanisms in Trading Environments with Moral Hazard. MIT Sloan Working Paper #4297-03, 2003.

[6] D. Fudenberg and D. Levine. Reputation and Equilibrium Selection in Games with a Patient Player. *Econometrica*, 57:759–778, 1989.

[7] R. Jurca and B. Faltings. An Incentive-Compatible Reputation Mechanism. In *Proceedings of the IEEE Conference on E-Commerce*, Newport Beach, CA, USA, 2003.

[8] 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.

[9] D. M. Kreps, P. Milgrom, J. Roberts, and R. Wilson. Rational Cooperation in the Finitely Repeated PRisoner's Dilemma. *Journal of Economic Theory*, 27:245–252, 1982.

[10] D. M. Kreps and R. Wilson. Reputation and Imperfect Information. *Journal of Economic Theory*, 27:253–279, 1982.

[11] P. Milgrom and J. Roberts. Predation, Reputation and Entry Deterrence. *J. Econ. Theory*, 27:280–312, 1982.

[12] N. Miller, P. Resnick, and R. Zeckhauser. Eliciting Honest Feedback in Electronic Markets. Working Paper, 2003.

[13] K. M. Schmidt. Reputation and Equilibrium Characterization in Repeated Games with Conflicting Interests. *Econometrica*, 61:325–351, 1993.

[14] B. Yu and M. Singh. An Evidential Model of Distributed Reputation Management. In *Proceedings of the AAMAS*, Bologna, Italy, 2002.