The Digitization of Word-of-Mouth: Promise and Challenges of Online Reputation Systems[1]

Chrysanthos Dellarocas
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139
dell@mit.edu

Abstract:

Online reputation mechanisms are emerging as a promising alternative to more traditional trust building mechanisms, such as branding and formal contracting, in settings where the latter may be ineffective or prohibitively expensive; a lot of electronic trading communities fall under these categories. Although a number of commercial websites already employ various forms of reputation mechanisms, rigorous research into their properties is still in its infancy. This fledgling field can benefit from past results in economics and game theory. Moreover, in order to translate the stylized results of game theory into concrete managerial guidance for implementing and participating in effective reputation mechanisms further advances are needed in a number of important areas: First, the design space of such mechanisms needs to be scoped and the effects of different design choices on performance need to be better understood. Second, the economic efficiency of various classes of reputation mechanisms needs to be quantified and compared to that of alternative mechanisms for building trust. Third, the robustness of those mechanisms against boundedly rational players, noisy ratings and strategic manipulation needs to be studied and improved. This paper surveys past results that have been derived in a variety of contexts, but which are relevant as a basis for building online reputation systems, presents two analytical models that illustrate the role of such systems in electronic markets and identifies opportunities for further MS/OR research in this fascinating area.

[1] Preliminary draft of December 10, 2001 for comments and discussion. Please do not circulate.

1. Introduction

Information asymmetries have a fundamental effect on the structure of markets. Akerlof (1970) has shown that, without counteracting institutions, the resulting adverse selection effects can lead to a reduction in the average quality of goods or even in the total dissolution of a market. A number of quality signaling and quality assurance mechanisms have, therefore, arisen to counteract the effects of quality uncertainty. Such mechanisms include branding and advertising, contractual guarantees (buyer's insurance, product warranties, money-back guarantees), government regulation (accreditation, licensing, legislation of minimum quality standards), consumer reports and informal, word-of-mouth networks. Information asymmetries also tend to make trading relationships "sticky" – uncertainty about quality of newcomers makes consumers reluctant to switch and induces producers to provide good service in order to maintain repeat business.

Far from eliminating information asymmetries, several properties of online interaction are, in fact, challenging the effectiveness of many traditional quality assurance mechanisms (Kollock 1999). Successful online marketplaces, such as eBay, are characterized by large numbers of small players, physically located around the world and often known to each other only via easily changeable virtual identities. Brand building is costly and impractical in such settings. Contractual guarantees are usually difficult or too costly to enforce due to the global nature of the market and the volatility of identities. For similar reasons, regulation is usually not effective, although things may change as governments become more proficient on how to legislate around the Internet. Finally, the large number of easily accessible players makes repeated interaction between a buyer and a seller less probable, thus reducing the incentives for a seller to behave well on the basis of creating a relationship and earning future business.

Fortunately, the situation is far from hopeless. One of the distinguishing features of the Web relative to previous media for mass communication is its bi-directionality: it's almost as easy to submit information to the Web, as it is to retrieve information from it. With proper design, the Web can be thought of as a low-cost medium for disseminating information to large user communities or as a low-cost medium for collecting and aggregating feedback from users.

An intriguing new family of electronic intermediaries are harnessing this property of the Web in order to provide a viable mechanism for quality assurance in cyberspace: *online reputation systems* (Resnick et. al. 2000), also known as recommender or feedback systems are attempting to engineer effective word-of-mouth networks in online environments for the purpose of signaling quality and inducing good behavior in the presence of information asymmetries.

Reputation systems collect feedback from members of an online community regarding past experiences with other members of that community. Submitted feedback is analyzed, aggregated with feedback received from other members and made publicly available to the community in the form of member *reputation profiles.*

Several examples of such mechanisms are already being used in a number of well-known online communities (Table 1). eBay, for example, relies on its feedback mechanism almost exclusively in order to both produce trust and induce good behavior on the part of its members. eBay buyers and sellers are encouraged to rate one another at the end of each transaction. A rating can be a designation of "praise", "complaint" or "neutral", together with a short text comment. eBay makes the cumulative ratings of its members, as well as all individual comments publicly available to every registered user. A growing body of empirical evidence seems to demonstrate that eBay's reputation system has managed to provide remarkable stability in an otherwise very risky trading environment (Bajari and Hortascu, 2000; Dewan and Hsu, 2001; Houser and Wooders, 2000; Lucking-Reiley et. al. 2000; Resnick and Zeckhauser, 2001).

The rising importance of online reputation systems for operators and participants of electronic marketplaces not only invites, but also necessitates rigorous research on their functioning and consequences. Are such mechanisms truly effective? Do they induce efficient market outcomes? To what extent can they be manipulated by strategic buyers and sellers? What is the best way to design them? How should buyers (and sellers) use the information provided by such mechanisms in their decision-making process? Under what circumstances are these mechanisms viable substitutes (or complements) of more traditional quality assurance mechanisms, such as branding and formal contracting? This is just a small subset of questions, which invite exciting and valuable management science research.

Web Site	Category	Summary of feedback mechanism	Type of feedback solicited	Type of reputation profiles
eBay	Online auction house	Buyers and sellers rate one another following transactions	Positive, negative or neutral rating plus comment	Sums of positive, negative and neutral ratings received during last 6 months
eLance	Professional services marketplace	Contractors rate their satisfaction with subcontractors	Numerical rating from 1-5 plus comment; subcontractor may post a response	Average of ratings received during last 6 months
Epinions	Online opinions forum	Users write reviews about products/services; other members rate the usefulness of reviews	Users rate multiple aspects of reviewed items from 1-5; readers rate reviews as "useful", "not useful", etc.	Averages of item ratings; % of readers who found a review "useful"
Google	Search engine	Search results are ordered based on how many sites contain links that point to them (Brin and Page, 1999)	How many links point to a page, how many links point to the pointing page, etc.	*No explicit reputation profiles are published; ordering acts as an implicit indicator of reputation*
Slashdot	Online discussion board	Postings are prioritized or filtered according to the ratings they receive from readers	Readers rate posted comments	

Table 1: Some noteworthy examples of online reputation mechanisms embedded in commercial web sites.

This paper aims to summarize past results and motivate further MS/OR research in the topic of online reputation system design and evaluation. Section 2 of the paper introduces the basic concepts and provides an overview of past work in game theory and economics on the topic or reputation. Section 3 illustrates the role of reputation mechanisms in market exchange settings by presenting and analyzing two simple models of such systems. Section 4 discusses several subtleties of reputational equilibria that need to be taken into account by designers of such systems. Section 5 presents a number of additional considerations that arise in the context of online interaction. Section 6 discusses how the set of insights provided by economics needs to be extended in order to produce theory-driven managerial guidance for implementing efficient and robust reputation systems in real-life settings. Finally, Section 7 summarizes the main conclusions and lists opportunities for future MS/OR research.

2. What is Reputation?

The concepts of reputation and word-of-mouth networks are as old as society itself. Long before the establishment of formal law and centralized systems of contract enforcement backed by the sovereign power of a state, ancient and medieval communities relied on word-of-mouth as the

primary enabler of economic and social activity (Benson, 1989; Greif, 1989, 1993; Milgrom, North and Weingast, 1990).

Whereas in "brick-and-mortar" societies word-of-mouth networks usually emerge naturally, online reputation mechanisms attempt to *artificially* induce the dynamics of those networks in cyberspace through the use of information technology. Their proper design and analysis requires a thorough understanding of how reputation effects arise naturally in social and economic situations. As a prerequisite to forming a discipline of reputation mechanism design, this and the following three sections provides an overview of results from game theory and economics that shed light to various aspects of this fascinating and surprisingly complex concept.

According to Wilson (1985), reputation is a characteristic or attribute ascribed to one person (organization, community, etc.) A by another person (or community) B. Operationally, this is often expressed as a prediction about A's likely future behavior ("A is likely to be courteous"). Nevertheless, it is primarily an empirical statement summarizing observations of A's past activity as perceived by B ("B says that A has been courteous to her in the past"). Its power in social interaction is based on the assumption that past behavior is indicative of future behavior.

Over the last twenty years, economists have tried to unravel this semantic tangle by the application of game theory. The key idea is that reputation is a *state variable* which

- Depends on observations of one's past actions by others
- Affects one's future payoffs.

As a concrete example, take Zagat's well-known numerical ratings of a restaurant's food quality. They are aggregate measures of ratings received by customers who already patronized that restaurant; therefore they summarize the past. Their power rests on the ability to influence a prospective customer's belief about the quality of food that she is likely to get in that restaurant, and therefore her decision of whether to patronize that restaurant in the future.

More generally, reputation is important in settings where

- There are multiple players
- At least one player has some private information that persists over time

- The informed player is likely to take several actions in sequence and is *unable to credibly commit* in advance to the sequence of actions he will take

In such settings, the future actions of uninformed players depend on their beliefs (probability assessments) about the informed player's private information. If there is an observable signal, which relates to the informed player's actions and can reveal something useful about her private information then that signal can update the uninformed players' beliefs and therefore their future behavior towards the informed player. Reputation thus creates an inter-temporal linkage along a sequence of otherwise independent situations: when selecting her next move, the informed player must take into account, not only her short-term payoff, but also the long-term consequences of her action based on what that action reveals about her hidden information to the other players.

The crucial factor that distinguishes reputation from other means of achieving social outcomes is the *absence of credible commitment*, often by all parties involved: If a seller can credibly pre-commit to provide good quality, or, conversely, if a buyer can credibly commit to send a seller to prison if he cheats her (which is pretty much the idea of signing a formal contract) then there is little need for the seller to maintain any reputation for honesty.

The absence of reliance on any form of commitment is reputation's biggest attraction: cooperation based on commitment usually requires costly (and often ineffective) enforcement technologies whereas cooperation based on reputation does not: reputation effects simply make cooperation everybody's best utility-maximizing response. On the other hand, this absence of commitment makes reputation effects rather fragile and their analysis significantly more convoluted. Furthermore, reputational equilibria do not come without cost: in most cases they result in inefficiencies whose cost needs to be carefully quantified and weighted against the benefits of not having to commit.

It's finally important to note that reputation effects do not necessarily benefit social welfare: depending on the detailed setting, they may result either in socially beneficial or in socially harmful outcomes. In some cases, reputation can induce cooperative behavior in settings where otherwise there wouldn't be any. In other cases, reputation allows one player to induce other players to behave in the way that best suits him, even though it may not be in their best interest to do so.

As an example of a socially beneficial outcome, consider the finitely repeated version of the symmetric prisoner's dilemma game (Kreps, Milgrom, Roberts and Wilson, 1982). It is well-known that in both the one-shot and the finitely-repeated version of this game, the dominant strategy of both players is to defect in every single round, thus realizing lower payoffs than if they had both cooperated.

The situation changes radically if there is some doubt, however small, in each player's mind about the other player's type. Suppose, for example, that each player believes that the other player may be a TIT-FOR-TAT type: a player who always cooperates as long as her opponent cooperates and always defects on the round immediately following a round where her opponent defected. Kreps et. al. show that, in that situation, it is in each player's long-term best interest to maintain their reputation for being TIT-FOR-TAT in the other player's mind by cooperating in all except perhaps a small number of rounds.

As an example of the second case, consider the chain store game (Kreps and Wilson, 1982). In the simplest version, an incumbent firm (the "store") plays in sequence against n opponents (the "entrants") the following simultaneous move stage game: each entrant chooses whether to open a store next to the incumbent's store or to stay out. If the entrant chooses to stay "out" he obtains a payoff of zero. If the entrant chooses to "enter" then his payoff depends on the incumbent's reaction: if the incumbent "fights" (e.g. engages in a price war) then the entrant gets –1. If the incumbent "accommodates" then the entrant gets +1. The incumbent's payoff is +2 if the entrant stays out, +1 if the entrant enters and the incumbent accommodates and –1 if the entrant enters and the incumbent fights. Each entrant maximizes its expected one-shot payoff, whereas the store maximizes the sum of its expected payoffs over the n-stage game. In the absence of information asymmetries it is obvious that the only equilibrium is (enter, accommodate) played at every stage: accommodate is a dominant strategy for the incumbent in every stage game and entrant is the entrant's best response.

Suppose now that we inject a bit of imperfect information into the game. More specifically, we assume that there is a nonzero probability in the mind of the entrant that the incumbent is a, so-called, "tough" type: he is determined to fight every entry irrespective of cost. In their seminal paper, Kreps and Wilson show how a "weak" (i.e. not "tough) incumbent can take advantage of

this uncertainty in order to maintain a reputation for toughness by "imitating" the behavior of a "tough" type: be prepared to fight entrants, even though this incurs short-term losses for the incumbent. Interestingly enough, knowing this, most entrants will not even attempt to enter. Reputation effects thus enable a "weak" incumbent to keep all (except possibly a small number of them) out of his home market.

The economics literature provides several other examples of situations where reputation effects arise. The interested reader is referred to Wilson (1985) for a literature survey. For the purposes of this paper, we will concentrate on the role of feedback mechanisms in market exchanges with asymmetric information. This is the topic of the next section.

3. Reputation Mechanisms in Market Exchanges

This section presents two models that illustrate the use of reputation mechanisms in addressing the adverse consequences of asymmetric information in electronic markets. When discussing the effects of asymmetric information, economists often make a distinction between adverse selection and moral hazard. *Adverse selection* describes situations where there is uncertainty about the other player's type (e.g. innate ability of a worker, reliability of a manufactured good, etc.). *Moral hazard* describes situations where there is uncertainty about the other player's actions (e.g. whether a hired worker will work as hard as promised, whether an eBay seller will ship back an item after receiving payment, etc.). Most real-life situations combine aspects of both adverse selection and moral hazard. However, to simplify the discussion, in the following sections we will illustrate the role of reputation in each case separately.

3.1 Reputation and adverse selection

Adverse selection effects arise when a market can potentially contain goods or services of different qualities. We are assuming that innate limitations restrict each producer to produce at a given quality. In such a market, the only hidden information is the type of the producer, i.e. the actual quality of the good. Consumers can only find out the true quality after purchasing the good. Producers then have no incentive to declare anything else except that their good is of highest quality. This signal would of course not be credible to consumers. In the absence of

counteracting institutions, Akerlof (1970) describes how, in such a market, all products except those of lowest quality would eventually be driven out.

We will illustrate Akerlof's argument in a market with two possible quality levels (L and H): Assume that consumers are willing to pay G_L for low quality products and $G_H > G_L$ for high quality products. Since the quality of a product is unknown, consumers would be willing to pay no more than $G_0 = p_L G_L + (1 - p_L) G_H$ where p_L is their prior belief about the probability of low quality. If G_0 is less than the cost of high quality products, only low quality producers would enter the market.

Introduction of a simple feedback mechanism can remedy the situation: We install a mechanism that encourages consumers, following each purchase, to vote their perception of a good's quality as either Low or High. We assume that quality perception is noisy and depends on such subjective factors as a consumer's mood, taste, etc. More specifically, the probability of a Low vote is a if a product is indeed of low quality and $0 < b < a$ if a product is of high quality. Our mechanism accumulates votes and disseminates the sum of low votes N_L and total votes N for each product in the form of a product reputation profile $\mathbf{R} = (N_L, N)$. Armed with this information, subsequent consumers can make better inferences about the expected fair value of products. More specifically, their reservation price is now a function of a product's reputation:

$$G(\mathbf{R}) = G(N_L, N) = p(L \mid \mathbf{R}) G_L + [1 - p(L \mid \mathbf{R})] G_H \qquad (1)$$

where $p(L \mid \mathbf{R})$ represents the updated belief that a product is of low quality, given its reputation profile. Using Bayes' law:

$$p(L \mid \mathbf{R}) = \frac{p(\mathbf{R} \mid L) p_L}{p(\mathbf{R} \mid L) p_L + p(\mathbf{R} \mid H)(1 - p_L)} = \frac{\binom{N}{N_L} a^{N_L} (1-a)^{N - N_L} p_L}{\binom{N}{N_L} a^{N_L} (1-a)^{N - N_L} p_L + \binom{N}{N_L} b^{N_L} (1-b)^{N - N_L} (1 - p_L)} \qquad (2)$$

As N grows, $N_L \to aN$ for low quality products and $N_L \to bN$ for high quality products. Therefore,

$$p(L \mid \mathbf{R}) \to \frac{e^{[a \ln a+(1-a)\ln(1-a)]N} p_L}{e^{[a \ln a+(1-a)\ln(1-a)]N} p_L + e^{[a \ln b+(1-a)\ln(1-b)]N}(1-p_L)} = \frac{\exp(-H_{aa}N)p_L}{\exp(-H_{aa}N)p_L + \exp(-H_{ab}N)(1-p_L)} \qquad (3a)$$

for low quality products, and

$$p(L \mid \mathbf{R}) \to \frac{e^{[b \ln a+(1-b)\ln(1-a)]N} p_L}{e^{[b \ln a+(1-b)\ln(1-a)]N} p_L + e^{[b \ln b+(1-b)\ln(1-b)]N}(1-p_L)} = \frac{\exp(-H_{ba}N)p_L}{\exp(-H_{ba}N)p_L + \exp(-H_{bb}N)(1-p_L)} \qquad (3b)$$

for high quality products, where $H_{xy} = -x \ln y - (1-x)\ln(1-y)$.

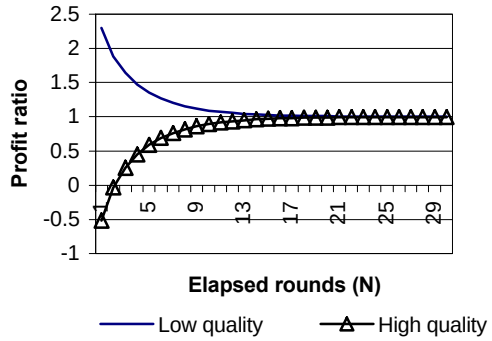
It is well known (see, for example, Savage 1971) that, for $0 \le x, y \le 1$ the expression H_{xy} is non-negative and is maximized when $x = y$. Therefore, if $0 < b < a$, $H_{aa} > H_{ab} \ge 0$ and $H_{bb} > H_{ba} \ge 0$. Substituting into (3a) and (3b), it is easy to see that, for all $0 < p_L < 1$, as $N \to \infty$, $p(L \mid \mathbf{R}) \to 1$ for low quality products and $p(L \mid \mathbf{R}) \to 0$ for high quality products. Substituting into (1) this means that, after a product has been on the market sufficiently long, $G(\mathbf{R}) \to G_L$ if the product is of low quality and $G(\mathbf{R}) \to G_H$ if the product is of high quality. Our simple feedback mechanism therefore succeeds in enabling consumers to eventually learn the true quality of products.

The average price of a product of low (high) quality after it has been on the market for *N* cycles is given by:

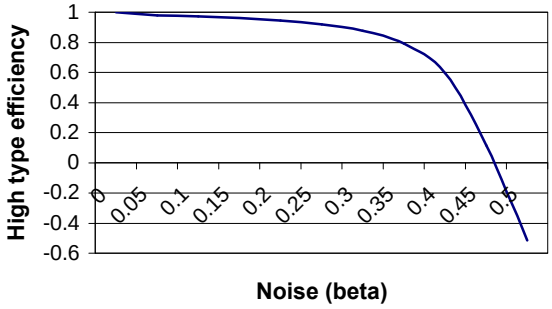
$$EG_L(N) = \sum_{n=0}^{N} p(\mathbf{R} = (n, N-n) \mid L) G(n, N-n) = \sum_{n=0}^{N} \binom{N}{n} a^n (1-a)^{N-n} G(n, N-n) \qquad (4a)$$

$$EG_H(N) = \sum_{n=0}^{N} p(\mathbf{R} = (n, N-n) \mid L) G(n, N-n) = \sum_{n=0}^{N} \binom{N}{n} b^n (1-b)^{N-n} G(n, N-n) \qquad (4b)$$

Figure 1(a) plots the profit ratios for low and high products (relative to the perfect information case) as a function of *N* and for indicative values of $a = 0.8$ and $b = 0.2$. We see that, during the initial phase, high quality producers initially receive lower profits than they deserve, while low quality producers realize higher profits than they deserve.



(a) $a = 0.8, b = 0.2, p_L = 0.7$, profit margin 30%



(b) $b = 1 - a$, $p_L = 0.7$, profit margin 30%

Figure 1: Profits and efficiency in a marketplace with adverse selection and a simple feedback mechanism.

The lifecycle expected payoff of a low quality seller is given by $V_L = \sum_{i=0}^{T} r^i (EG_L(i) - c_L)$, and similarly for a high quality seller. Finally, one plausible measure of the *stakeholder efficiency* of the mechanism for a particular player type is the ratio of her expected lifecycle payoff in this market over the equivalent ratio in a market with perfect information. It is easy to see that the simple reputation mechanism presented in this example results in low efficiency for high quality producers (see Figure 1(b)). Efficiency declines with noise (the higher the b and the lower the a). If there is sufficient noise and profit margins are low enough, efficiency may even be negative. In that case it is not worthwhile for high quality sellers to build a reputation in this market, therefore they will stay out and the "lemons" effect will still prevail.

Feedback mechanisms whose goal is to estimate a "static" property of a good or service are the simplest ones to design and analyze[2]. Nevertheless, the proper design and evaluation of even this simple set of mechanisms involves some subtle issues. One important set of issues comes from the fact that, in many settings, perception of quality is subjective and therefore, the expected valuation of a product depends, not only on the ratings that the product has received in the past, but also on the person who is inquiring. These, and other related concerns, have motivated research in the area of personalization, which is covered in detail in another article of this special issue (Add Reference).

[2] This special class of reputation mechanisms is often referred to as *recommender systems* (Schafer et. al. 2001).

3.2 Reputation and moral hazard

Almost every commercial transaction in which one party moves (sends payment, ships the product, etc.) before the other party completes its part of the transaction is exposed to moral hazard. In such situations, the player who moves second has a choice between at least two moves (e.g. cooperate, cheat) with the following property: the move which is most beneficial for the second player (cheat) is least beneficial for the first player and vice versa. Furthermore, the first player has no direct power over the second player's choice of move.

The situation can be represented as a game with sequential moves and prisoner's dilemma-like payoffs. It is well known that in a single-shot version of such an exchange game the dominant strategy is for both sides to renege, i.e. no transaction will take place.

The situation improves if the game is infinitely repeated between one buyer and one seller. This models long-term relationships between two trading partners. In such cases, if the payoffs from cooperation are high enough for both parties, each party can induce good behavior from the other party by threatening termination of the relationship if the other party ever defects.

When the number of buyers increases, the problem of maintaining cooperation is exacerbated. Assume that a single long-term seller is facing *N* buyers in round-robin fashion. In that case, if buyers do not communicate with one another, punishment from defection with buyer *i* will not take effect until the seller meets again with the same buyer. In the meantime, the seller can reap additional unfair payoffs by cheating other buyers. As the number of buyers increases, the temptation for cheating will be stronger. If the seller's discount rate r is strictly less than 1, then, there is always a number $N_0(r)$ with the property that, if the number of buyers transacting with that seller grows above N_0, it will be rational for the seller to cheat (Croson, 1996).

The limiting case is a situation where a single long-term seller is facing a sequence of one-shot buyers. Each buyer buys only once from the given seller no matter what the outcome of the transaction. This situation is equivalent to the single-shot prisoner's dilemma game: no transaction will ever take place. It is arguable that several online environments have properties that approximate such a setting. On eBay, for example, there are close to 4 million members, each of which can be both a buyer and a seller. Given the diversity of products being traded and

the fact that products are searched by keyword and not by seller, repeat business by the same customer is highly unlikely. It is in such settings that reputation mechanisms are indispensable

We will illustrate their functioning by describing a simple model that abstracts some of the features of eBay's feedback mechanism. Consider a marketplace with a single long-term seller and a sequence of homogeneous one-shot buyers. The seller sells a single type of product of known quality and unit cost 1. Buyers are willing to pay $(1+m)$ for a unit of the product and therefore leave the seller with a profit m. At each stage of the game, the seller moves first posting a price for the product. Buyers then decide whether to buy or not buy. If they decide to buy, they send payment to the seller. Finally, after receiving payment, the seller decides whether to cooperate, that is, send the product, thus realizing a short-term payoff of m, or cheat, in which case the payoff is $(1+m)$. The marketplace has a feedback mechanism, which encourages buyers to rate a transaction as positive or negative. If a seller cheats, he will get a negative vote with probability 1. Because consumer assessments are noisy, even if he does not cheat, he may still get a negative vote with probability $0 \le p \le 0.5$ [3]. The feedback mechanism publishes the history (sequence) of positive and negative votes that the seller has received so far.

In this game, the seller's objective is to maximize the present value of his profits over the entire span of the game, while the buyer's objective is to maximize her short-term (stage game) surplus. The key assumption is that, at the beginning of the game, buyers have a belief (probability) h_0 that the seller may be of "honest" (irrational) type: a type that never cheats, no matter what. Subsequent buyers update this belief based on the seller's feedback profile. That belief serves as the seller's *reputation* for honesty.

In the following analysis we will index stages by the number of *remaining* rounds until the end of the game. That is, stage *–1* is the last stage of the game, stage *–2* the next-to-last, etc.

Since quality is fixed, the only uncertainty in this game is about whether the seller will cheat. Therefore, at stage $-i$ buyers would be willing to pay no more than:

$$G_{-i} = [h_{-i} + (1 - h_{-i})b_{-i}](1+m) \qquad (5)$$

[3] In the highly unlikely case where $p > 0.5$ then positive votes play the role of negative votes and the following analysis produces the same results if we substitute "negative vote" with "positive vote" throughout.

where b_{-i} is the *buyer's* belief about the probability that a "strategic" (rational) seller will cooperate in round -*i*. Assuming that the buyer agrees to pay G_{-i} and sends the money to the seller, the seller must then determine how to respond (cooperate or cheat). In the most general case, the seller's response at stage $-i$ is characterized by a probability s_{-i} of cooperation. At equilibrium, the buyer's belief b^*_{-i} must equal the seller's optimal response s^*_{-i} at each stage.

To solve for the sequential equilibrium of the finite-horizon version of this game we will first solve for the sequential equilibrium of the one-period game, then that of the two-period game, and proceed by induction to solve for the game with *N* periods.

It is easy to determine the sequential equilibrium of a single play of this game: Strategic sellers will always cheat and buyers will expect them to do so. In other words, $b^*_{-1} = s^*_{-1} = 0$ and $G_{-1} = h_{-1}(1+m)$.

Suppose now that there are two periods remaining in the game. The seller's objective function is:

$$V_{-2}(h_{-2}) = \max_{b_{-2}} \max_{s_{-2}(b_{-2})} (G_{-2} - s_{-2}) + r[p^+(s_{-2})V_{-1}(h^+_{-1}) + p^-(s_{-2})V_{-1}(h^-_{-1})] \quad (6)$$

subject to the constraints $0 \le b_{-2} \le 1,\ 0 \le s_{-2} \le 1$ and $b^*_{-2}(h_{-2}) = s^*_{-2}(b^*_{-2}(h_{-2}))$.

In the above equation, G_{-2} is the price buyers are willing to pay in period -2, s_{-2} is the short-term cost of cooperating with probability s_{-2} (unit cost is 1), r is the discount factor, $p^{\pm}(s_{-2})$ is the probability that the seller will get a positive or negative rating in stage -2 given his response, and h^+_{-1}, h^-_{-1} are subsequent buyers' updated beliefs that a seller may be honest given that the seller received a positive or negative rating respectively in round -2. From the preceding discussion, it is easy to see that:

$$G_{-2} = [h_{-2} + (1-h_{-2})b_{-2}](1+m) \quad (7a)$$

$$V_{-1}(h^{\pm}_{-1}) = G_{-1}(h^{\pm}_{-1}) = h^{\pm}_{-1}(1+m) \quad (7b)$$

In round -2, a strategic seller will get a negative rating with probability 1 if he cheats and with probability p if he cooperates. Since s_{-2} is the probability of cooperating at round -2:

$$p^{-}(s_{-2}) = (1 - s_{-2}) + s_{-2}p = 1 - s_{-2}(1-p) \quad (8a)$$

$$p^{+}(s_{-2}) = s_{-2}(1-p) \quad (8b)$$

For honest sellers, the corresponding probabilities are p and $1-p$ respectively.

Since subsequent buyers do not directly observe s_{-i} (this is *exactly* the source of moral hazard in this game!), they update their beliefs about a seller's probability of being honest given the seller's rating by using b_{-i} in place of s_{-i}. Using Bayes' rule:

$$h^{+}_{-i+1}(h_{-i}) = \frac{(1-p)h_{-i}}{(1-p)h_{-i} + p^{+}(b_{-i})(1-h_{-i})} = \frac{(1-p)h_{-i}}{(1-p)h_{-i} + b_{-i}(1-p)(1-h_{-i})} = \frac{h_{-i}}{h_{-i} + b_{-i}(1-h_{-i})} \quad (9a)$$

and

$$h^{-}_{-i+1}(h_{-i}) = \frac{ph_{-i}}{ph_{-i} + p^{-}(b_{-i})(1-h_{-i})} = \frac{ph_{-i}}{ph_{-i} + [(1-b_{-i}) + b_{-i}p](1-h_{-i})} \quad (9b)$$

3.2.1 Noise-free ratings

The situation is easiest to analyze when $p = 0$ (noise-free ratings). In this case, a seller who does not cheat will never get a negative rating. Conversely, if a seller gets a negative rating then this completely reveals to buyers that the seller is strategic. If buyers are certain that they are dealing with a strategic seller then the game reverts to a finitely repeated prisoner's dilemma game: In that case, no buyer will buy from the seller in any subsequent round. For simplicity, we drop stage indices for b, h. For $p = 0$ equation (6) becomes:

$$V_{-2}(h) = \max_{b} \max_{s(b)} (h + (1-h)b)(1+m) - s + rsV_{-1}\left(\frac{h}{h + b(1-h)}\right) \quad (10)$$

subject to the constraints $0 \le s \le 1, \ \ 0 \le b \le 1$ and $s^* = b^*$.

For a given b, the seller's best response s maximizes $W(b) = -s + rsV_{-1}\left(\frac{h}{h + b(1-h)}\right)$. $W(b)$ is linear in *s*. More specifically: $\frac{\partial W(b)}{\partial s} = -1 + rV_{-1}\left(\frac{h}{h + b(1-h)}\right)$. Furthermore, for a given h, $V_{-1}\left(\frac{h}{h + b(1-h)}\right) = \frac{h}{h + b(1-h)}(1+m)$ ranges from $h(1+m)$ to $(1+m)$, attains its maximum value for $b = 0$ and declines as b grows.

The resulting equilibrium depends on the prior belief h that a seller is honest, and the value of the remaining payoff in round –1.We can distinguish three cases:

a. If $r(1+m)<1$ then $rV_{-1}<1$ and $\frac{\partial W(b)}{\partial s}<0$ for all b and all h, which means that $W(b)$ is maximized at $s=0$ (always cheat). In words, if the discounted remaining payoff in round –1 is less than the cost of cooperating in round –2, no matter what the seller's reputation in round –2, then it is optimal for strategic sellers to cheat. In this case, the only equilibrium is for strategic sellers to cheat in round *-2* and for the buyers to expect them to do so, which means that sellers cannot set second round prices any higher than $h(1+m)$. Since no buyer will purchase in the final round, the two stage discounted payoff in this case is $V_{-2}=G_{-2}=h(1+m)$. The reputation mechanism fails to induce the players to cooperate.

b. If $r(1+m)\geq 1$ and $r(1+m)h\geq 1$ then $rV_{-1}\geq 1$ and $\frac{\partial W(b)}{\partial s}\geq 0$ for all b, so $W(b)$ is maximized at $s=1$ (never cheat). Therefore, in cases where both the profit margin and the reputation of a seller for honesty are high enough, a strategic seller will not cheat in round -2 and can credibly signal that to buyers by pricing his product using $b^*=s^*=1$. The two stage discounted seller payoff is $V_{-2}=(1+m)-1+rV_{-1}=m+rh(1+m)$. The reputation mechanism thus succeeds in inducing cooperation in round -2.

c. Finally, if $r(1+m)\geq 1$ but $r(1+m)h<1$, then $\frac{\partial W(1)}{\partial s}<0$: if buyers believe that a strategic seller will never cheat $(b=1)$, it is optimal for the seller to select $s=0$ (always cheat); certain cooperation is not an equilibrium. Furthermore, $\frac{\partial W(0)}{\partial s}>0$: if buyers believe that the seller will always cheat, it is optimal for the seller to never cheat; certain defection is not an equilibrium either. The only equilibrium corresponds to a mixed strategy: buyers expect a strategic seller to cooperate with probability $s^*=b^*<1$, such that the seller is indifferent between cheating and cooperation. In other words, b^* is selected so that $\frac{\partial W(b^*)}{\partial s}=0\Rightarrow b^*=\frac{[r(1+m)-1]h}{1-h}$. If we substitute this value of $b^*=s^*$ into (10) the two stage discounted seller payoff is $V_{-2}=rh(1+m)^2$. In this case the reputation mechanism is not

entirely successful in inducing cooperation, but still manages to reduce the probability of cheating. The relative effectiveness of the mechanism depends on the value of $r(1+m)h$: the higher the profit margin and/or the higher the h, the higher the equilibrium probability b^* that the seller will cooperate and the higher the price that the seller can charge. Note that $h=0 \Rightarrow b^*=0$: if buyers are certain that the seller is strategic, the seller will cheat, therefore the buyers will not buy; the situation reverts to a single-shot prisoner's dilemma game. The existence of even a small belief of honesty is essential for reputation effects to occur.

With three periods remaining, in the noise-free case the seller's objective is to maximize:

$$V_{-3}(h_{-3}) = \max_{b_{-3}} \max_{s_{-3}(b_{-3})} (h_{-3} + (1-h_{-3})b_{-3})(1+m) - s_{-3} + rs_{-3}V_{-2}\left(\frac{h_{-3}}{h_{-3} + b_{-3}(1-h_{-3})}\right) \qquad (11)$$

subject to the constraints $0 \le s_{-3} \le 1,\ \ 0 \le b_{-3} \le 1$ and $s^*_{-3}(b^*_{-3}) = b^*_{-3}$.

For a given $b_{-3} \equiv b$ the seller's optimal response $s_{-3} \equiv s$ maximizes:

$$W(b) = -s + rsV_{-2}(h') \quad \text{where } h'(h,b) = \frac{h}{h + b(1-h)} \ge h \qquad (12)$$

Once again, we distinguish three cases:

a. If $r(1+m) < 1$ then $rV_{-2}(h') = rh'(1+m) < 1$ and $\frac{\partial W(b)}{\partial s} = -1 + rh'(1+m) < 0$ for all h, b. Therefore, the only equilibrium is for a strategic seller to cheat and for the buyers to expect him to do so.

b. If $r(1+m) \ge 1$ and $rV_{-2}(h) \ge 1$ then, since $h' \ge h$, $\frac{\partial W(b)}{\partial s} = -1 + rV_{-2}(h') \ge -1 + rV_{-2}(h) \ge 0$ for all b. It is an equilibrium for the buyers to expect the seller to cooperate and for the seller to do so. In that case $h' = h$, that is, the seller's subsequent reputation remains unchanged between rounds. Note that, if it optimal for the seller to cooperate in round –2 (which implies $r(1+m)h' \ge 1$) then $rV_{-2} = r[m + rh'(1+m)] \ge r(m+1) \ge 1$ so the seller will always cooperate in round –3 as well.

c. If $r(1+m) \geq 1$ but $rV_{-2}(h) < 1$, the seller will cooperate with probability $b^* < 1$, such that $\frac{\partial W(b^*)}{\partial s} = 0$. This situation only arises if $r(1+m)h' < 1$, in which case $V_{-2}(h') = rh'(1+m)^2$. and gives $s^* = b^* = \frac{[r^2(1+m)^2 - 1]h}{1-h}$.

Note that, for $r(1+m) \geq 1$ the equilibrium probability of cooperation in round –3 is always greater than or equal to the equilibrium probability of cooperation in round –2.

Continuing by induction, we can now visualize the evolution of the game if *N* stages remain:

If, initially, $h_0 = 0$ then buyers rightfully expect all sellers to cheat and therefore they will not buy at any price. If, on the other hand, buyers have a, however small, non-zero prior belief h_0 that the seller may be honest, then the evolution of the game depends on the profit margin:

- If $r(1+m) < 1$, strategic sellers cheat during the first period, receive a negative rating and exit. Their long-term payoffs are not large enough to offset their short-term gains from cheating and reputation effects cannot help in this case. Anticipating this, buyers are not willing to pay more than $h_0(1+m)$ on any transaction.
- If $r(1+m) \geq 1$, then,
    - As long as the remaining horizon is long enough so that the remaining discounted payoff $rV_{-i+1}(h_0) \geq 1$, strategic sellers cooperate and buyers expect sellers with no negative ratings to do so. Therefore, buyers are willing to pay the full information price $(1+m)$. Furthermore, as long as a seller keeps getting positive ratings he maintains his reputation for honesty h_0 unchanged from one round to the next (his reputation goes to 0 if he gets even a single negative rating though).
    - As the end of the game approaches, there comes a stage when the remaining discounted payoff if sellers maintain their reputation becomes less than the unit cost. From then on, sellers begin to "milk" their reputation by cheating with nonzero probability. The idea is that, at each round the seller randomizes between cheating and cooperation. Sellers who cheat receive negative votes and exit. Sellers who cooperate proceed to the next round

with *higher* reputation (this is why it is worthwhile to the seller to gamble between cheating and not cheating). The probability of cheating increases as sellers get closer to the end of the game and becomes equal to 1 in the very last round. Anticipating this behavior, during the endgame buyers will only accept to buy at increasingly lower prices.

3.2.2 Noisy ratings

The introduction of even a small amount of noise (probability p of receiving an "unfair" negative rating even if the seller cooperates) radically changes the behavior of the system.

Generalizing (6), we can write the seller's objective function as:

$$V_{-i}(h_{-i}) = \max_{b_{-i}} \max_{s_{-i}(b_{-i})} (G_{-i} - s_{-i}) + r[s_{-i}(1-p)V_{-i+1}(h^{+}_{-i+1}) + [1 - s_{-i}(1-p)]V_{-i+1}(h^{-}_{-i+1})] \quad (13)$$

subject to the constraints $0 \le s_{-i} \le 1,\ \ 0 \le b_{-i} \le 1$ and $s^*_{-i}(b^*_{-i}) = b^*_{-i}$.

For a given b_{-i}, the seller's optimal response s_{-i} maximizes:

$$W_{-i}(b_{-i}) = -s_{-i} + r[s_{-i}(1-p)V_{-i+1}(h^{+}_{-i+1}) + [1 - s_{-i}(1-p)]V_{-i+1}(h^{-}_{-i+1})] \quad (14)$$

which gives $$\frac{\partial W_{-i}(b_{-i})}{\partial s_{-i}} = -1 + r(1-p)[V_{-i+1}(h^{+}_{-i+1}) - V_{-i+1}(h^{-}_{-i+1})] \quad (15)$$

From equations (9a) and (9b) it is easy to see that if $p > 0$, for $b_{-i} = 1$, $h^{+}_{-i+1} = h^{-}_{-i+1} = h_{-i}$ and $\frac{\partial W_{-i}(1)}{\partial s_{-i}} = -1$, which means that $W_{-i}(1)$ is maximized for $s_{-i} = 0$. In words, if at *any* stage, buyers believe that a strategic seller will always cooperate, it is optimal for him to always cheat. Therefore, this cannot be an equilibrium. This result is important because it means that, with even a small amount of noise, this class of reputation mechanisms can *never* 100% eliminate the possibility of cheating.

We will now study the behavior of the system when the remaining horizon is very long. In this case, $V_{-i}(h) = V_{-i+1}(h)$. Dropping the stage indices, equations (9ab),(13),(15) can be rewritten as:

$$V(h) = \max_{b(h)} \max_{s(b(h))} [h + (1-h)b(h)](1+m) - s + r\Big(s(1-p)V(h^+) + [1 - s(1-p)]V(h^-)\Big) \quad (16)$$

$$\frac{\partial W(b)}{\partial s} = -1 + r(1-p)[V(h^+) - V(h^-)] \quad (17)$$

$$h^+(h,b) = \frac{h}{h + b(1-h)} \quad (18)$$

$$h^-(h,b) = \frac{ph}{ph + [(1-b) + bp](1-h)} \quad (19)$$

For $b = 0$, $h^+(h,0) = 1, h^-(h,0) = \frac{ph}{ph + (1-h)}$ and $\frac{\partial W(0)}{\partial s} = -1 + r(1-p)[V(1) - V(h^-(h,0))]$. As h grows, $h^-(h,0)$ approaches 1. Therefore, there exists some $\bar{h}$ (typically very close to 1) with the property that $\frac{\partial W(0)}{\partial s} < 0$ for $h > \bar{h}$. As before, $\frac{\partial W(0)}{\partial s} < 0$ implies $s^* = b^* = 0$. In words, if a strategic player's reputation grows above a threshold $\bar{h}$, his optimal response is to cheat always and buyers should expect him to do so. Cheating will result in the seller's reputation falling in the next round. Therefore, no strategic seller can consistently maintain a reputation above $\bar{h}$.

Since we have ruled out the possibility of an equilibrium where $s^* = b^* = 1$, for $h \le \bar{h}$ the only possible equilibria correspond to mixed strategies $s^*(h) = b^*(h) < 1$, such that $\frac{\partial W(b^*)}{\partial s} = 0$. From (16), (17), the condition $\frac{\partial W(b^*)}{\partial s} = 0$ corresponds to:

$$V(h^+) - V(h^-) = \frac{1}{r(1-p)} = \text{constant}(h) \quad (20)$$

$$V(h) = [h + (1-h)b^*(h)](1+m) + rV(h^-) \quad (21)$$

This is a continuous state, discrete time dynamic programming problem and its exact solution can only be found numerically.

Figure 2 shows an approximate solution for $q = 0.01, m = 0.3$ and $r = 0.9999$. We see that, for strategic sellers, the probability of cooperation stays at around 80% but falls for very large and very small reputations. The same figure shows the prices that buyers are willing to pay as a function of the seller's reputation for honesty relative to the price in a market with perfect information. Prices are always lower than in the perfect information case. This is an efficiency loss due to information asymmetries.

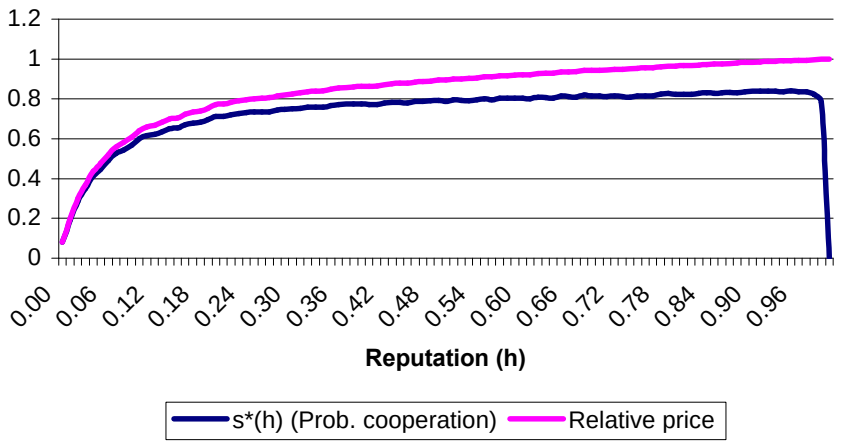


Figure 2: Probability of collaboration and relative prices in a market with moral hazard and a noisy feedback mechanism

4. The Fragility of Reputation

The models of the preceding section make it apparent that reputation is a powerful, yet fragile concept. Its effectiveness usually varies with time. Furthermore, it is often dependent on several other environmental parameters in subtle and unexpected ways. Finally, reputational equilibria do incur a cost, in the form of reduced efficiency. This cost must be compared with the cost of achieving similar outcomes through other methods (such as formal contracting). Designers of successful online reputation mechanisms must be intimately familiar with these considerations and this section provides a brief overview of related results and open questions.

4.1 Dynamic Considerations

Reputation is a highly time sensitive property. Upon entry of a new player it needs to be built. Once built, it needs to be maintained. Finally, in games with finite time horizons (in reality, all games) players have an opportunity to realize some additional gains during the last few stages by "milking" their reputation before exiting.

Initial phase: Building a reputation

Bootstrapping a reputation mechanism is not trivial. At the beginning of a new seller's lifecycle, buyers have no information about her type. In order for a reputation to be built in the first place, some buyers should be willing to transact with the seller at some price. Furthermore, these courageous buyers should disseminate valuable feedback to the rest of the community. It is not obvious why they would be willing to do so. More generally, it is not obvious why members of an online community would contribute feedback to a reputation mechanism. In fact, basic economic principles predict that evaluations, which are public goods, are likely to be underprovided. Avery, Resnick and Zeckhauser (1999) study this problem and present pricing and subsidy mechanisms that operate through a computerized market and induce the efficient provision of evaluations

During the initial phase, it is not uncommon that some players realize lower, or even negative profits, while the community "learns" their true type. In those cases, the players will only enter the market if that initial phase cost is offset by the present value of the gains from reputation in the later part of the game. As we have shown in the model of Section 3.1, in noisy environments this is not always a trivial consideration.

Another consideration is whether the incentive effects of reputation to induce "good" behavior work right away. The answer here depends on the specifics of the situation. In most cases, reputation effects begin to work immediately and in fact, are strongest during the initial phase, when players must work hard to establish a reputation. Holmstrom (1982) discusses an interesting model of reputational considerations in the context of an agent's "career" concerns: suppose that wages are a function of an employee's innate ability for a task. Employers cannot directly observe an employee's ability, however, they can keep track of the average value of her past task outputs. Outputs depend both on ability and labor. The employee's objective is to maximize her lifecycle's wages while minimizing the labor she has to put in. At equilibrium, this provides incentives to the employee to work hard right from the beginning of her career. In fact these incentives are strongest at the very beginning of her career when observations are most informative.

In contrast, Diamond's (1989) analysis of reputation formation in debt markets presents a setting where reputation effects do not work right away. In Diamond's model there are three types of borrowers: safe borrowers, who always select safe projects (i.e. projects with zero probability of default), risky borrowers, who always select risky projects (i.e. projects with higher returns if successful but with nonzero probability of default) and strategic borrowers who will select the type of project that maximizes their long term expected payoff. The objective of lenders is to maximize their long term return by offering competitive interest rates, while at the same time being able to distinguish profitable from unprofitable borrowers. Lenders do not observe a borrower's choice of projects, but they do have access to her history of defaults. This is the state variable that plays the role of reputation. In Diamond's model, if lenders believe that the initial fraction of risky borrowers is significant, then, despite the reputation mechanism, at the beginning of the game interest rates will be so high that strategic players have an incentive to select risky projects. Some of them will default and will exit the game. Others will prove lucky and will begin to be considered as safe players. It is only after lucky strategic players have already acquired some initial reputation (and therefore begin to receive lower interest rates) that it becomes optimal for them to begin "masquerading" as safe players by consciously choosing safe projects in order to maintain their good reputation.

Steady state: Maintaining a reputation

Reputational games are ideally characterized by a steady-state equilibrium where the state variable interpreted as "reputation" reaches a value that remains stable over time (barring small stochastic fluctuations). The most important question is whether such a steady-state equilibrium exists in the first place. The alternative would be a situation in which a seller finds it optimal to oscillate, repeatedly building up a reputation and then milking it. If oscillations were optimal, the predictive value of reputation would be significantly diminished. Shapiro (1982) has studied this problem and has come up with a set of concavity conditions that are sufficient for the existence of steady-state quality levels.

Given that "reputation" is usually based on imperfect observations of one's actions, another important consideration in this stage is whether the objective of maintaining one's "reputation" (state variable) at steady-state levels is sufficient to continuously induce players to behave well.

Once again, the answer depends on the exact details of what state variable is interpreted as reputation and how it is updated over time. It is interesting to note that in the case most commonly encountered in practice, that is, basing reputation on the average of all past observations of one's behavior, the answer to the above question is a resounding no. The longer a player stays in the game, the smaller the incremental effect of her current actions on her *lifecycle* average. Therefore, contrary to intuition, long-lived players have less incentives to behave well than short-lived players! This phenomenon has been discovered by Holmstrom (1982) in his analysis of career concerns: his model predicts that, at equilibrium, employees have weaker incentives to work hard in later stages of their career than earlier on; once they have established a good reputation, their reputation, expressed as an average of their outputs over their *entire* career, stays with them even if their later outputs decline. The implication of this phenomenon for online reputation mechanisms is clear: replace simple averages with weighted averages that give proportionally more importance to recent feedback.

Endgame: Milking one's reputation

One of the limitations of using reputation to induce good behavior is that reputational effects only work if the remaining horizon of a seller is long enough. As we have seen in Section 3.2 as soon as the seller approaches exit, the cost of maintaining a reputation exceeds the remaining benefits and the seller finds it optimal to begin cheating with increasingly higher probability.

One possible solution is to assign some post-mortem value to reputation, so that players find it optimal to maintain it through the end of the game. For example, reputations can be viewed as assets that can be bought and sold in a market for reputations (Tadelis 1999, 2000).

Another possible solution is to levy a security deposit to all newcomers in a marketplace that will only be refunded upon exit if the exiting player's reputation is good enough. This is an especially interesting idea in online communities, because it also helps alleviate some of the problems associated with "cheap pseudonyms", discussed in more detail in Section 5.

4.2 Dependence on Profit Margins

Reputation effects induce players to forego short-term losses in order to realize larger long-term gains. Obviously, they can only be effective if the latter exceed the former. An interesting result

of Section 3.2 is that, in order for reputation to work, not only the remaining horizon must be long enough, but also the profit *per transaction* must exceed a threshold. Similar results have been obtained by Klein and Leffler (1981) and Shapiro (1983). Shapiro calls this phenomenon the premium for reputation. This result can have at least two potential interpretations: (a) reputation mechanisms are not effective in highly competitive markets or (b) in markets where trust in based on reputation, prices tend to be higher than in markets with perfect information. Both of these conjectures are very interesting and require further theoretical and empirical investigation.

4.3 Noise and imperfect correlation between observable feedback and actual behavior

In most real-life settings, reputation is based on imperfect observations of a signal. As we saw in Section 3 imperfect observations, or noise, can often qualitatively change the resulting equilibria and incur efficiency losses. Additional issues arise when the signal used as feedback is imperfectly correlated with the behavior that it purports to reveal/control. In those cases, players might be able to strategically modify their behavior in order to be able to generate a "good quality" signal but still hide or misrepresent their true types. One striking example from a real-life setting comes from the study by Dranove et. al (2000) of the effects of health care report cards (public disclosure of patient health outcomes at the level of the individual physician or hospital). The purpose of health care report cards was to inform the public about the performance of hospitals and physicians and provide incentives for low performing health providers to perform better. Initial studies of New York's report cards in the 1990s concluded that the implementation of the mechanism reduced mortality (as reported on the score card). Deeper scrutiny revealed, however, that, following the implementation of the system, hospitals changed their patient selection procedures and accepted fewer high-risk patients. Therefore, they were able to improve their score card rating without necessarily improving their performance.

5. Additional considerations in online environments

5.1 "Cheap Pseudonyms"

In most online communities the notion of identity is rather fickle. Members usually know each other only through pseudonyms, which are difficult to trace to a "physical" person. Furthermore,

members can usually change pseudonyms very easily and with minimal cost, hence the term cheap pseudonyms, coined by Friedman and Resnick (1999).

In such environments long-term players have one additional strategy available to them: they can disappear at any stage and reappear under a new identity (and a clean record). In some settings this would allow seller's to realize unfair profits by periodically entering, building a good reputation, milking it, exiting and reentering under a new identity. It is therefore an important consideration when designing online reputation mechanisms.

Friedman and Resnick discuss two classes of approaches to this issue: Either make it more difficult to change online identities, or structure the community in such a way so that exit and re-entry with a new identity becomes unprofitable. The first approach makes use of various cryptographic authentication technologies and is outside the scope of this paper. The second approach is based on imposing an upfront cost to each new entrant, such that the benefits of "milking" one's reputation are exceeded by the cost of subsequent re-entry. This cost can be an explicit entrance fee or an implicit cost of having to go through a reputation building (or *"dues paying"*) stage with low or negative profits.

Generally speaking, if cheap pseudonyms are a concern in the community where the mechanism will be embedded, then, in addition to all other requirements, the mechanism must satisfy a *no milking and re-entry* property: This property specifies that it should not be optimal for a player to deviate from the steady state by milking his reputation, exiting and re-entering under a new identity. More formally, the discounted present value of the payoff of the optimal path of milking, followed by reentry, should be less than the discounted present value of the payoff gained by remaining at steady state during the same period. Currently, the conditions under which a reputation mechanism satisfies this property are an open area for research.

5.2 Strategic manipulation of reputation mechanisms

As online reputation mechanisms begin to exercise greater influence on decision-making, the incentive for strategically manipulating them becomes correspondingly stronger. The low cost of submitting online feedback coupled with the relative anonymity of the rater makes such

manipulation a real problem that needs to be studied and addressed before such mechanisms can achieve widespread adoption.

A number of commercial reputation mechanisms (for example, Epinions.com) have attempted to address this problem through the concept of "rate the rater": members can rate how useful other members' feedback has been to them. Whereas this technique is somewhat effective for reducing noise, it is not effective for reducing strategic manipulation. Determined manipulators can manipulate the "rate the rater" ratings as much as they can manipulate the ratings themselves.

So far there are relatively few results in this important area. Dellarocas (2000a, 2000b) has pointed out a number of manipulation scenarios and has proposed a number of decision-theoretic immunization mechanisms that reduce the effect of such attacks if the fraction of unfair raters is reasonably small (up to 20-30% of the total population). Mayzlin (2000) has analyzed the impact of strategic manipulation of online bulletin boards.

One of the issues that complicate addressing of strategic manipulation of reputation mechanisms is a relative asymmetry that exists in most of today's reputation mechanisms between the incentives of the ratee and those of the rater. Whereas bad behavior by the ratee results in a negative rating by the rater, which, in turn, reduces the future payoffs of the ratee, in most systems, dishonest behavior by a rater does not incur any future consequences for the rater other than, at best, having his ratings ignored. Even in environments where both parties have the right to rate one another, there are still significant asymmetries: for example, on eBay, where sellers can also rate buyers, a bad rating received by a buyer carries very small "punch" because a seller does not have the right to decline bids from "bad" buyers, whereas a buyer can certainly refrain from bidding on a "bad" seller's auctions.

Further progress in this area requires mechanisms that impose some type of cost for dishonest ratings (or, conversely, reward raters for honest behavior). But this, once again, is subject to moral hazard issues: since it is not currently possible to read people's minds, honesty cannot be directly observed. In some cases, dishonest ratings can be identified as statistical outliers. In most cases, however, ratings involve some degree of subjectivity, in which case it is difficult to tell whether an outlier represents a dishonest rating or a genuinely different belief. Prelec (1987)

proposes a number of game-theoretic mechanisms that induce truth telling if the raters cannot form coalitions. This is clearly an area where further research is urgently needed.

6. From models to systems: Towards an MS/OR discipline of reputation system design

The preceding sections presented several examples from the literature of economics that illustrate the value of game theory in providing an initial characterization of the effects of reputation mechanisms. Since these mechanisms are emerging as an attractive alternative to more traditional trust building institutions in a variety of trading communities (ranging from electronic marketplaces to inter-organizational supply chains), their proper design and implementation is becoming a topic of interest for the management scientist.

In contrast to economics, whose emphasis is on stylized analysis and general insights, the objective of MS/OR is to provide theory-driven guidance that will assist electronic market operators and participants to make optimal choices in real-life settings. Therefore, in order to evolve the initial set of insights provided by economics into a MS/OR discipline of online reputation system *design*, further advances in at least three areas are required: First, the design space of such mechanisms needs to be scoped and the effects of different design choices on performance need to be better understood. Second, the economic efficiency of various classes of reputation mechanisms needs to be quantified and compared to that of other alternative trust building mechanisms. Third, the robustness of those mechanisms against boundedly rational players, noisy ratings and strategic manipulation needs to be studied and improved.

6.1 Scoping the design space of reputation mechanisms

In a sense, the problem of reputation mechanism design is the inverse of that of game theory: Rather than fix a game and look for the set of equilibrium outcomes, the objective of a reputation mechanism designer is to fix the set of desirable outcomes and then design an online environment that implements a game, which in turn yields those outcomes as equilibria. It is therefore important for the designer to identify the relevant design dimensions of such an environment, as well as the configurations of design parameters that result in well-performing mechanisms.

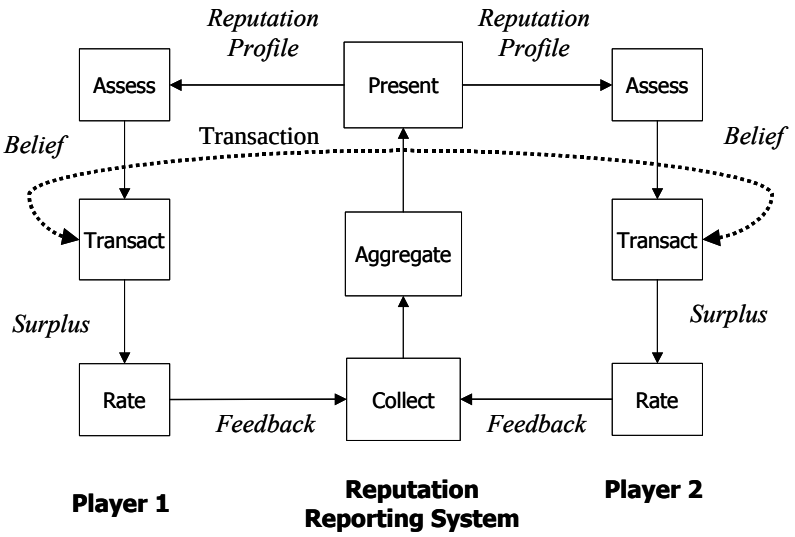


Figure 3: Simplified architecture of a reputation mechanism.

Tangible design dimensions

The most "tangible" component of an online reputation mechanism (Figure 3) is a reputation reporting system: an information system that collects, aggregates and disseminates "reputational information". The most important design dimensions associated with such a system include:

- *Type of feedback to be solicited.* Examples include binary feedback (i.e. "good"/"bad", "positive"/"negative", etc.), ratings on a scale (e.g. from 1 to 5), comparison feedback (did you like product A better than B?), etc.
- *Feedback aggregation procedure.* Examples include the use of simple statistical measures (such as average, weighted average, moving average, median, etc.), collaborative filtering (Bresee et. al., 1998) and a number of more complicated, but typically ad-hoc, algorithms proposed by computer scientists (Zacharia et. al. 2000; Sabater and Sierra, 2001; Yu and Singh, 2001).
- *Form of reputation profiles.* Examples include fractional numbers, star ratings, histograms of ratings, etc. Some systems, such as Google and Slashdot, do not explicitly publish reputation profiles, but rather use internally computed profiles as the basis of performing some action (e.g. ordering of search results). In those cases the outcome of the action relative to an item (e.g. the ranking of a search result) acts as an implicit indication of that item's reputation.

- *Profile initialization rules.* Most systems start off new members with zero reputation. Other plausible alternatives include initializing newcomers with average reputation, or even allowing new members to purchase units of reputation upon entry.
- *Participation rules.* This dimension specifies the restrictions on who and when is allowed to post ratings. Are all members allowed to post feedback about any other member, do ratings need to refer to specific transactions performed through the system, etc.
- *Identity rules.* Are raters/ratees known through their real identities or through pseudonyms? How easy is it to change pseudonyms? Does the system identify members who have changed identity in a special way?

So far there is little work in exploring the properties of specific subsets of this design space. For example, Dellarocas (2000a, 2000b) has identified certain weaknesses of the currently prevalent ratings aggregation method based on average and has proposed alternative ratings aggregation algorithms with better robustness properties. He also argued that a form of controlled anonymity in which the center discloses the reputation of members but not their identity could shield members against badmouthing attacks. In another paper, Dellarocas (2001) explores aspects of the effectiveness of binary reputation mechanisms. Further work is needed to identify what other design dimensions are important which design configurations work best in what settings.

Intangible design dimensions

One of the complications associated with trying to artificially create word-of-mouth networks is that the effectiveness and efficiency of such networks depends on a number of intangible dimensions that are not under the direct control of the mechanism designer. The most important intangible dimensions are the profile assessment function and the rating function.

The *profile assessment function* specifies how the (tangible) information contained in a feedback profile gets translated into a (intangible) belief and a decision (e.g. to transact or not) in the reader's mind. Since this process takes place inside somebody's mind, the mechanism designer has not direct control over it. Of course, according to the principles of game theory, we "know" that rational decision-makers have a well-specified method for updating their beliefs: bayesian belief updating. It is equally well known, however, that humans are rarely acting in accordance with Bayes rule (El-Gamal and Grether, 1995; see Conlisk, 1996 for an extensive bibliography

on bounded rationality). When modeling the performance of reputation mechanisms in realistic settings, therefore, it is worthwhile to also consider behavioral models of decision-making (see Payne, Johnson and Bettman, 1992 for a survey and bibliography). Further empirical and experimental research has an important role to play in this area.

The *rating function* maps the outcome of a transaction into one of the possible ratings in the feedback scale (or to a decision to not rate at all). Initial evidence demonstrates that the exact form of that function can have significant impact on the effectiveness and efficiency of the reputation mechanism. For example, in his analysis of the economic efficiency of binary reputation mechanisms, Dellarocas (2001) has shown that more lenient ratings (coupled with stricter interpretation of feedback profiles) result in higher efficiency for the seller. Intuitively, the rating function ought to be a function of the buyer's satisfaction with the transaction. The marketing literature provides a wealth of theories and experimental results related to consumer satisfaction in "brick-and-mortar" settings (e.g. Oliver, 1980, 1996). These results need to be re-validated and extended to reflect buyer behavior in the online domain.

6.2 Quantifying the efficiency of reputation mechanisms

As we discussed in Section 4, reputational equilibria almost always incur some loss of efficiency relative to equivalent outcomes in environments with perfect information. This efficiency loss has multiple sources, including the learning that takes places during the initial phase, reputation milking during the endgame, steady-state inefficiencies due to mixed equilibria, noisy ratings, etc. Furthermore, efficiency losses are usually different for different player types. When considering the use of a reputation mechanism in an electronic market, the market designer must carefully quantify this cost and compare it to the overhead of other, more traditional trust building institutions, such as formal contracting.

In Section 3.1 we proposed stakeholder efficiency as one plausible efficiency metric in such. Another plausible metric is the social efficiency, defined as the ratio of the expected social welfare in a marketplace using the mechanism over the expected social welfare in a similar market with perfect information. The right efficiency metric will depend on the strategic objectives of the reputation mechanism.

6.3 Determining and improving robustness properties

The examples of Sections 3 and 4 have demonstrated that reputational equilibria are quite fragile. For example, the introduction of even a small amount of noise in Section 3.2 has qualitatively changed the nature of the resulting equilibria. Given this, once an initial mechanism has been designed and analyzed, it is important to study how robust it is to changes in the underlying assumptions. The most important considerations here are to study how changes in the assumptions about the intangible dimensions of the system (profile interpretation function, rating function, initial beliefs, profit margins, etc.) affect the equilibria of the system. Another consideration of particular relevance in online settings is to study robustness against noisy ratings, as well as strategic manipulation, including unfair ratings and strategic identity changes.

7. Conclusions

Online reputation mechanisms harness the remarkable ability of the Web to, not only disseminate, but also collect and aggregate information from large user communities at very low cost, in order to artificially construct large-scale word-of-mouth networks. Such networks have historically proven to be valid quality assurance mechanisms in settings where information asymmetries can adversely impact the functioning of a community and where formal contracting is unavailable, unenforceable or prohibitively expensive. A lot of online communities currently fall into these categories. Therefore, online reputation mechanisms are emerging as a promising alternative to more traditional trust building mechanisms in the digital economy.

The design of such mechanisms can greatly benefit from the insights produced by more than twenty years of economics and game theory research on the topic of reputation. To that end, this paper presented an overview of relevant results and introduced two illustrative models that demonstrate the role of simple feedback mechanisms in addressing adverse selection and moral hazard issues in exchange settings. The most important conclusion drawn from this survey is that reputation is a powerful but subtle and complicated concept. Its power to induce cooperation without the need for costly and inefficient enforcement institutions is the basis of its appeal. On the other hand, its effectiveness is often ephemeral and depends on a number of additional tangible and intangible environmental parameters. Finally, social outcomes that rely on

reputation do incur efficiency losses, which need to be quantified and compared with the overhead associated with alternative methods of achieving similar outcomes.

In order to translate the stylized results of economics into concrete guidance for implementing and participating in effective reputation mechanisms further advances are needed in a number of important areas. The following list contains what the author considers to be the most important open areas of MS/OR research in reputation mechanism design:

- Scope and explore the design space and limitations of reputation mechanisms. Understand what set of design parameters work best in what settings. Develop formal models of those systems in both monopolistic and competitive settings.
- Compare the relative efficiency of reputation mechanisms to that of other mechanisms for dealing with information asymmetries (such as formal contracting and brand-name building) and develop theory-driven guidelines for deciding which mechanisms to use when.
- Develop effective solutions to the problems of identity change and strategic manipulation.
- Quantify the value of reputation mechanisms to the various stakeholders and propose viable business models for reputation mechanisms.

Furthermore, the design of those systems will benefit from further experimental and empirical research that sheds more light into buyer and seller behavioral models vis-à-vis such mechanisms.

Applications such as online reputation mechanisms, which attempt to artificially engineer the dynamics of heretofore naturally emergent social systems, are opening a new exciting chapter on the frontiers of information technology. New methodologies, that combine insights from management science, economics, sociology, psychology and computer science are needed in order to fully understand and design them. The author is looking forward to further research in this useful and exciting area.

References

Akerlof, G. (1970) The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, pp. 488-500.

Avery, C., Resnick, P. and Zeckhauser, R. (1999) The Market for Evaluations. *American Economics Review*, 89, 3, pp. 564-584.

Bajari, P. and A. Hortascu (2000). Winner's Curse, Reserve Prices and Endogenous Entry: Empirical Insights From eBay Auctions. Working paper, December 2000.

Benson, Bruce (1989). The Spontaneous Evolution of Commercial Law. *Southern Economic Journal* 55, Jan.: 644-61. Reprinted in *Reputation: Studies in the Voluntary Elicitation of Good Conduct*, edited by Daniel B. Klein, 165-89. Ann Arbor: University of Michigan Press, 1997.

Bresee, J.S., Heckerman, D., and Kadie, C. (1998) Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pp. 43-52, San Francisco, July 24-26, 1998.

Brin S. and Page L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the Seventh World Wide Web Conference* (WWW7), Brisbane, also in a special issue of the journal *Computer Networks and ISDN Systems*, Volume 30, issues 1-7.

Conlisk, J. (1996) Why Bounded Rationality? *Journal of Economic Literature*, 34, 2, pp. 669-700.

Croson, D. (1996). A New Role for Middlemen: Centralizing Reputation in Repeated Exchange. Chapter 1 in "Improving Allocative Efficiency: Middlemen, Management Principles and Channel Pricing". Unpublished Ph.D. Dissertation, Program in Business Economics; Harvard University.

Dellarocas, C. (2000a). Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. *Proceedings of the 2nd ACM Conference on Electronic Commerce*, Minneapolis, MN, October 17-20, 2000

Dellarocas, C. (2000b). The Design of Reliable Trust Management Systems for Online Trading Communities. Working Paper, Center of eBusiness, MIT Sloan School of Management.

Dellarocas. C. (2001). Analyzing the economic efficiency of eBay-like online reputation reporting mechanisms. *Proceedings of the 3rd ACM Conference on Electronic Commerce*, Tampa, FL, October 14-16, 2001.

Dewan, S. and Hsu, V. (2001) Trust in Electronic Markets: Price Discovery in Generalist Versus Specialty Online Auctions. Working Paper. January 31, 2001.

Diamond, D. (1989). Reputation Acquisition in Debt Markets. *Journal of Political Economy* 97, 4, pp. 828-862.

Dranove, D., Kessler, D., McClellan, M., and Sattherthwaite, M. (2000) Is More Information Better? The Effects of 'Report Cards' on Health Care Providers. Working Paper.

El-Gamal, M.A. and Grether, D. M. (1995). Are People Bayesian? Uncovering Behavioral Strategies. *Journal of the American Statistical Association*, 90, 432, pp. 1137-45.

Friedman, E. and Resnick, P. (2001). The Social Cost of Cheap Pseudonyms. *Journal of Economics and Management Strategy* **10**(1).

Greif, A. (1989). Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders. *Journal of Economic History* 49, December 1989, pp. 857-82.

Greif, A. (1993). Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition. *American Economic Review* 83, June 1993, pp. 525-548.

Greif, A., Paul Milgrom, and Barry R. Weingast. (1994). Coordination, Commitment, and Enforcement: The Case of the Merchant Guild. *Journal of Political Economy* 102: 745-76.

Holmstrom, B. (1982). Managerial Incentive Problems: A Dynamic Perspective. In *Essays in Economics and Management in Honour of Lars Wahlbeck*. Helsinki: Swedish School of Economics, pp. 209-230.

Houser, D. and Wooders, J. (2000) Reputation in Auctions: Theory and Evidence from eBay. Working Paper, University of Arizona, 2000.

Klein, B. and Leffler, K. (1981) The Role of Market Forces in Assuring Contractual Performance. *Journal of Political Economy*. 89, 4, pp. 615-641.

Kollock, P. (1999) The Production of Trust in Online Markets. In *Advances in Group Processes* (Vol. 16), eds. E.J. Lawler, M. Macy, S. Thyne, and H.A. Walker, Greenwich, CT: JAI Press.

Kreps, D., Milgrom, P., Roberts, J and Wilson, R. (1982). Rational Cooperation in the Finitely Repeated Prisoners' Dilemma. *Journal of Economic Theory*, 27, pp. 245-52.

Kreps, D. and Wilson, R. (1982). Reputation and Imperfect Information. *Journal of Economic Theory*, 27, pp. 253-79.

Lucking-Reiley, D., Bryan, D., Prasad, N. and Reeves, D. (2000). Pennies from eBay: The Determinants of Price in Online Auctions, Working Paper, Vanderbilt University, 2000.

Mayzlin, D. (2000). Promotional Chat on the Internet. Working Paper, Yale University, 2000.

Milgrom, P. R., Douglass C. N., and B. R. Weingast. (1990). The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs. *Economics and Politics* 2: 1-23.

Oliver, R.L. (1980) A Cognitive model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of Marketing Research*, 17 (November), pp. 460-469.

Oliver, R. (1996) Satisfaction: A Behavioral Perspective on the Consumer. New York: McGraw-Hill.

Payne, J., Bettman, J.R. and Johnson, E.J. (1992). Behavioral Decision Research: A Constructive Processing Perspective. *Annual Rev. Psych*., 43, pp. 87-131.

Prelec, D. (1987). Introspection and Communication. A Game-Theoretic Approach. Harvard University, April 1987.

Resnick, P., Zeckhauser, R., Friedman, E., Kuwabara, K. (2000) Reputation Systems. *Communications of the ACM*, Vol. 43, (12), December 2000, pp. 45-48.

Resnick, P. and Zeckhauser, R. (2001) Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System. Working Paper for the NBER workshop on empirical studies of electronic commerce. January 2001.

Sabater, J. and Sierra, C. (2001). REGRET: A reputation model for gregarious societies. Proc. *4th Workshop on Deception, Fraud and Trust in Agent Societies*, Montreal, Canada, May 2001.

Savage, L.J. (1971). Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association*, Vol. 66, No. 336. (Dec., 1971), pp. 783-801.

Schafer, J.B., Konstan, J., and Riedl, J., (2001) Electronic Commerce Recommender Applications. *Journal of Data Mining and Knowledge Discovery*. January, 2001.

Shapiro, C. (1982) Consumer Information, Product Quality, and Seller Reputation. *Bell Journal of Economics* 13 (1), pp 20-35, Spring 1982.

Shapiro, C. (1983). Premiums for High Quality Products as Returns to Reputations. *The Quarterly Journal of Economics*, November 1983, pp. 659-679.

Tadelis, S. (1999). What's in a Name? Reputation as a Tradeable Asset. *The American Economic Review* **89**(3): 548-563.

Tadelis, S. (2000). The Market for Reputations as an Incentive Mechanism, Stanford University: Working Paper. June 2000

Wilson, Robert (1985). Reputations in Games and Markets. In *Game-Theoretic Models of Bargaining*, edited by Alvin Roth, Cambridge University Press, pp. 27-62.

Yu, B. and Singh, M.P. (2001). Towards a Probabilistic Model of Distributed Reputation Management. Proceedings of the *4th Workshop on Deception, Fraud and Trust in Agent Societies*, Montreal, Canada, May 2001.

Zacharia, G., A. Moukas, et al. (2000). Collaborative Reputation Mechanisms in Electronic Marketplaces. *Decision Support Systems* **29**(4): 371-388.