Analyzing the economic efficiency of eBay-like online reputation reporting mechanisms[¶]

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

Abstract

This paper introduces a model for analyzing marketplaces, such as eBay, which rely on binary reputation mechanisms for quality signaling and quality control. In our model sellers keep their actual quality private and choose what quality to advertise. The reputation mechanism is primarily used to induce sellers to advertise truthfully. Buyers base their ratings on the difference between expected and actual quality. Furthermore, raters are lenient and do not post negative ratings unless transactions end up exceptionally bad. It is shown that, in such a setting, the fairness of the market outcome is determined by the relationship between rating leniency and corresponding strictness when assessing a seller's feedback profile. If buyers judge sellers too strictly (relative to how leniently they rate) then, at steady state, sellers will be forced to understate their true quality. On the other hand, if buyers judge too leniently then sellers can get away with consistently overstating their true quality. An optimal judgment rule, which results in outcomes where, at steady state, buyers accurately predict the true quality of sellers, is theoretically possible to derive for all leniency levels. Furthermore, if buyers judge sellers using that rule, then the more lenient buyers are when rating sellers, the more likely it is that sellers will find it optimal to settle down to steady-state quality levels, as opposed to oscillating between good quality and bad quality. However, it is argued that this optimal rule depends on several parameters, which are difficult to estimate from the information that eBay currently makes available to its members. It is therefore questionable to what extent unsophisticated buyers are currently using eBay feedback information in an optimal way.

[¶] This is work in progress. Please do not quote or distribute without the author's permission. Suggestions and comments very welcome. Please address correspondence to dell@mit.edu

1. Introduction

Online reputation reporting systems are emerging as an important quality signaling and quality control mechanism in online trading communities (Kollock 1999; Resnick et. al. 2000). Reputation systems collect feedback from members of an online community regarding past transactions with other members of that community. Feedback is analyzed, aggregated and made publicly available to the community in the form of member *feedback profiles*. If one accepts that past behavior is a relatively reliable predictor of future behavior, then these profiles can act as a powerful quality signaling and quality control mechanism, essentially acting as the digital equivalent of a member's *reputation*.

eBay relies on its reputation 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 sums of praise, complaint and neutral ratings submitted for each member, as well as all individual comments, publicly available to all its users. Anecdotal and empirical results seem to demonstrate that eBay's reputation system has managed to provide remarkable stability in an otherwise very risky trading environment (Dewan and Hsu 2001; Resnick and Zeckhauser 2001).

The rising practical importance of online reputation systems not only invites but rather necessitates rigorous research on their functioning and consequences. Are such mechanisms truly reliable? Do they promote efficient market outcomes? To what extent are they manipulable 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? This is just a small subset of unanswered questions, which invite exciting and valuable research.

The study of reputation as a mechanism for inducing good behavior in markets with asymmetric information is certainly not new. Several economists have published important works analyzing its properties (Rogerson, 1983; Schmalensee, 1978; Shapiro, 1982; Smallwood and Conlisk, 1979; Wilson, 1985, just to name a few).

Nevertheless, although past work in economics has studied some of the overall effects of reputation, it has paid very little attention to the analysis of *specific* mechanisms for forming and communicating reputation, in part because in traditional brick and mortar societies such mechanisms are largely informal (they are often referred to as "word-of-mouth advertising") and defy detailed modeling. The few published results focusing on the effects of specific properties of reputation mechanisms clearly make the point that such properties can have significant effects on the market outcome. For example, Rogerson (1983) shows that reputation based on subjective binary ratings (e.g. good/bad, praise/complaint) creates an externality, which affects the entire market. Shapiro (1982) shows that, unless the mechanism by which reputation is formed satisfies certain properties, sellers may find it optimal to continuously oscillate in quality, periodically building good reputation and subsequently milking it.

On the other hand, the design and implementation of online reputation systems has so far been the research domain of computer scientists (see Bresee et. al., 1998; Sarwar et. al., 2000; Schafer et. al. 2001 for overviews of past work). The emphasis of past work in the area has been on developing algorithms and systems for collecting, aggregating and extracting useful information from sets of user ratings, drawing from work in information retrieval, data mining and collaborative filtering. The analysis and evaluation of the proposed algorithms is typically done in terms of computational complexity and statistical metrics, such as their running time, memory requirements, average recall and precision, average bias, etc.

We believe that there is a need for work that bridges the two disciplines: research, which takes into account the algorithmic details of specific reputation systems but also models how these systems are embedded inside trading communities and investigates their effectiveness and impact, not only in terms of computational and statistical properties, but rather in terms of their overall impact in the efficiency of the market and the welfare of the various classes of market participants. Given that reputation systems were conceived in order to assist choice in environments of imperfect information, their impact in those latter market dimensions should be the ultimate determinant of success of any new proposed new algorithm and system.

This paper contributes in this direction by proposing a model for analyzing the economic efficiency of binary reputation systems, such as the one used by eBay.

Section 2 introduces the model and its underlying assumptions. We assume that buyer satisfaction on eBay is a function of the difference between the advertised and true quality of an item. In such a setting, the reputation mechanism is primarily used to induce sellers to advertise truthfully. Section 3 formalizes this intuition into a number of properties that eBay-like reputation mechanisms should satisfy, in order to be considered *well functioning*.

Section 4 applies our model in order to determine under what circumstances such mechanisms can indeed be well functioning. It is shown that the fairness of the market outcome is determined by the relationship between rating leniency and judgment strictness when assessing a seller's feedback profile. An optimal judgment rule, which results in outcomes where, at steady state, buyers accurately predict the true quality of sellers, is theoretically possible to derive for all leniency levels. A rather surprising conclusion of our analysis is that, if buyers use this optimal judgment rule, then the more lenient buyers are when rating sellers, the more likely it is that sellers will find it optimal to settle down to steady-state quality levels, as opposed to oscillating between good quality and bad quality. In that sense, rater leniency *benefits* the overall stability of the system. However, it is argued that this optimal rule depends on several parameters, which are difficult to estimate from the feedback information that eBay currently makes available to its members. It is therefore questionable to what extent unsophisticated buyers are currently using eBay feedback information in an optimal way.

Section 5 considers the implications of relaxing some of the simplifying assumptions on which our analysis is based. Finally, Section 6 summarizes the contributions and conclusions of the paper.

2. A model of reputation-mediated marketplaces with binary feedback

This section introduces a model for analyzing marketplaces, which rely exclusively on a binary reputation mechanism for quality signaling and quality control. A *binary reputation mechanism* is a mechanism where raters are given the opportunity to rate past transactions using one of two values, commonly interpreted as “positive” (i.e. satisfactory) and “negative” (i.e. unsatisfactory, problematic). Our intention is to use this model in order to study the economic impact of reputation mechanisms similar to the one used by eBay (see Resnick and Zeckhauser 2001 for a detailed description)[1].

In our model, qualities are non-negative real-valued quantities, which subsume aspects of both product quality and service quality. We assume that each seller produces items, whose *real quality* q_r is unknown to buyers and can only be determined with accuracy after consumption. We further assume that all buyers prefer higher quality to lower quality, although they might differ in the extent to which they are prepared to pay for an extra unit of quality. Finally, we assume that although the real quality of items is not communicated to buyers, sellers do inform buyers by advertising. On eBay, advertising corresponds to the seller-supplied description, which accompanies all items. The *advertised quality* q_a of an item is completely controlled by the seller (i.e. there is no validation of any kind by a third party) and may or may not correspond to its real quality.

Sellers aims to maximize the present value of their payoff function $\boldsymbol{p}(x,q_r,q_a)=G(x,q_r,q_a)-c(x,q_r)$ where *x* is the volume of sales, *G*(.) is the gross revenue function and *c*(.) is the cost function. We assume that $\partial c/\partial q_r \geq 0$ and $\partial \boldsymbol{p}/\partial q_a \geq 0$ for all sellers.

Under the above assumptions, sellers have an incentive to over-advertise quality. The market would then degenerate to a “market for lemons” (Akerlof 1970). In order to avoid this from happening, buyers are given the option to rate each transaction using a “positive” or “negative” rating. A reputation system, operated by a trustworthy third party, accumulates all ratings into a feedback profile $\mathbf{R}=(\Sigma_+,\Sigma_-,\Sigma_{\text{no rating}})$ for each seller, where Σ_+ is the sum of all positive ratings received for that seller during the most recent time window, Σ_- is the sum of all negative ratings received during the same period and

[1] In addition to “positive” (praise) and “negative” (complaint) ratings, eBay’s reputation mechanism also supports “neutral” ratings (which, however, are rarely used in actual practice). As will become apparent below, our model subsumes raters who would submit “neutral” ratings on eBay into the set of raters who

$\Sigma_{\text{no rating}}$ is the number of transactions for which no rating was submitted[2]. Time windowing is used in order to address the possibility that sellers may improve or deteriorate their behavior over time. For example, on eBay, feedback profiles display the sums of ratings received during the past 6 months only.

Buyer utility from purchase of a single item is modeled by $U = q \cdot q - p$, where p is the price, q is a buyer's quality sensitivity and q is the level of quality perceived by the buyer *after* consumption. When considering a purchase, buyers combine all the information that is available to them, i.e. an item's advertised quality and a seller's feedback profile, in order to form a subjective assessment of an item's *estimated quality* q_e, where:

$$q_e = f(q_a, \mathbf{R}) \tag{1}$$

Armed with knowledge of prices and estimated qualities, buyers proceed to purchase one of the available items, presumably the one which maximizes their expected utility $U_e = q \cdot q_e - p$. Following a purchase, buyers observe the item's perceived quality $q = q_r + e$, where e is a normally distributed error term with standard deviation s. The introduction of an error term is intended to collectively model a number of phenomena, which occur in actual practice. For example:

- buyers may misinterpret a seller's advertised quality (this should be modeled as $q_e = q_a + e$, however, our analysis is identical if we add the error term to q instead)
- sellers may exhibit small variations in actual quality from one transaction to another
- buyers may have small differences in their perception of quality based, say, on their moods that day
- some aspects of perceived quality depend on factors beyond a seller's control (e.g. post-office delays)

Finally, buyers decide whether to rate a transaction as well as what rating to give. Our model assumes that ratings are a function of a buyer's satisfaction relative to her expectations. We define a buyer's satisfaction from a given transaction to be the difference between perceived and expected utility. That is, $S = U - U_e = q \cdot (q_r - q_e + e)$. Under the above assumptions, S is a normally distributed random variable with mean $q \cdot (q_r - q_e)$ and standard deviation $q \cdot s$.

One interesting property of eBay, which has been reported on several empirical studies, is that most buyers give very few negative ratings to sellers. Resnick and Zeckhauser (2001) report that less than 0.5% of buyers and 1.2% of sellers post neutral or negative feedback about their trading partners (see Figure 1). They tentatively conclude that raters either rate generously or prefer to refrain from rating at all after bad experiences.

[2] eBay does *not* currently publish $\Sigma_{\text{no rating}}$. The results of this paper make a strong case that they should.

	Buyer of Seller		Seller of Buyer	
	Frequency	Percent	Frequency	Percent
Negative	111	0.3	353	1.0
Neutral	62	0.2	60	0.2
Positive	18,569	51.2	21,560	59.5
None	17,491	48.3	14,260	39.4
Total	36,233		36,233	

Figure 1: Frequencies of feedback in Resnick and Zeckhauser's data set.

Resnick and Zeckhauser (2001) refer to this phenomenon as a "high courtesy equilibrium" and offer several speculative explanations: eBay allows reciprocal ratings (that is, sellers also rate buyers) and buyers are often afraid that posting a negative rating for a seller will lead to retaliatory bad ratings, harassing emails etc. eBay does not provide mechanisms to prevent or assist such situations. Furthermore, it has been reported that sellers often communicate with buyers via email and negotiate settlements to transaction problems, while explicitly pleading with them to not post negative ratings. Finally, eBay has created a "culture of praise", where the vast majority of ratings and comments are extremely positive. In such a setting, most buyers feel a moral obligation to conform to the prevailing social norms and be nice and relatively forgiving to their trading partners.

Our model uses a rating function *r*(*S*), which attempts to model the above empirically observed behavior. More specifically, we are assuming that buyers rate a transaction as "positive" if their actual utility from the transaction exceeds their expected utility (i.e. if *S>0*). On the other hand, buyers only rate a transaction as negative if their actual utility falls short of their expected utility by more than a *leniency factor* l, that is, if $S < -l$. Finally, for transactions, which end up being "slightly bad but not too bad" (i.e. where $-l < S \leq 0$), we are assuming that buyers prefer to simply refrain from rating at all[3]. To summarize:

$$r(S) = \begin{cases} \text{"+"} & \text{if S} > 0 \\ \text{"–"} & \text{if } S \leq -l \\ \text{no rating} & \text{if } -l < S \leq 0 \end{cases} \qquad (2)$$

where $S = U - U_e = q \cdot (q_r - q_e + e),\ e \sim N(0, s)$,

To simplify the initial analysis, we are making the assumption that q, s and l are constant across the entire population of buyers and sellers. In Section 5, we will relax those assumptions and study how they impact the results derived in Sections 3 and 4.

3. Well functioning reputation mechanisms

The following sections will use the model developed in Section 2, in order to explore under what circumstances binary reputation mechanisms can be *well functioning*. Before

[3] On eBay, some buyers would post a "neutral" rating in this case.

doing that, however, in this section we will discuss what it means for a reputation mechanism to be well functioning in marketplaces with private quality information. We *define* a well-functioning reputation mechanism to be one, which satisfies the following two properties:

> **WF1:** If there exists an equilibrium of prices and qualities under perfect information (i.e. in settings where $q_e = q_a = q_r$) then, in environments where q_r is private to sellers, the existence of the reputation mechanism makes it optimal for sellers to settle down to a steady-state pair of real and advertised qualities, rather than to oscillate, successively building up and milking their reputation.
>
> **WF2:** Assuming WF1 holds, under all steady-state seller strategies (q_r, q_a) the quality of sellers as estimated by buyers *before* transactions take place, is equal to their true quality (i.e. $q_e = q_r$).

Before we proceed, let us justify the above definition by providing a brief rationale for the desirability of properties WF1 and WF2.

First, the value of reputation mechanisms in general relies on the assumption that past behavior is a reliable predictor of future behavior (Wilson 1985). If oscillations were optimal, the predictive value of cumulative functions of past ratings, such as Σ_+, Σ_-, would be greatly diminished. In environments where the primary (or only) mechanism for certifying and controlling seller quality is based on reputation, it is, therefore, desirable that sellers find it optimal to settle down to a steady-state behavior rather than to oscillate.

Second, according to our model, buyers make purchase decisions based on knowledge of prices and estimated qualities. If it were possible for sellers to settle down into a steady-state strategy that would consistently deceive buyers into estimating $q_e > q_r$, then this would allow sellers to earn additional profits at the expense of buyers. In the presence of competitive marketplaces, buyers would then eventually leave the marketplace in favor of other markets with better information. On the other hand, if, under all possible steady-state seller strategies the effect of the reputation mechanism was such, so that buyers estimated $q_e < q_r$, then the opposite effect would take place: buyers would realize extra surplus at the expense of sellers. Once again, we would then expect that sellers would desert the marketplace in favor of other, more transparent markets. The only fair steady-state strategy, therefore, is one where $q_e = q_r$.

4. Can binary reputation mechanisms be well functioning?

This section will demonstrate that, given a rating function, which has the general form given by (2), whether an eBay-like binary reputation system satisfies property WF2 depends on the relationship between (2) and the quality estimation function $q_e = f(q_a, \mathbf{R})$. Furthermore, we will show that, if buyers are lenient enough when they rate and correspondingly strict when they judge seller profiles, sellers will find it optimal to settle down to steady-state true and advertised quality levels if such an equilibrium exists under perfect information.

4.1 Estimated vs. real qualities in steady state

Let us first focus our attention on the circumstances under which a binary reputation mechanism satisfies condition WF2. We are assuming that WF1 holds. Therefore, there exists at least one steady-state strategy (q_r, q_a) for each seller[4]. A steady-state strategy is a strategy that optimizes a seller's payoff function, while at the same time resulting in an estimated quality $q_e = f(q_a, \mathbf{R})$, which is stable over time. Denote $q_e = q_r + x$, where x is the *deception factor*, that is, the distortion between estimated and real quality at steady state. If $x > 0$ then buyers overestimate a seller's true quality, whereas if $x < 0$ then buyers underestimate true quality. Let *N* be the total number of sales transactions of a given seller in the most recent time window. It is easy to see that $N = \Sigma_+ + \Sigma_- + \Sigma_{no\ rating}$. Assuming that buyers rate according to (2), for large *N* at steady state the following will hold:

$$\begin{aligned} \Sigma_+ &= N \cdot \Pr ob[S > 0] = N \cdot \Phi[(q_r - q_e)/s\ ] = N \cdot \Phi[-x/s\ ] \\ \Sigma_- &= N \cdot \Pr ob[S \le -l\,] = N \cdot \Phi[(q_e - q_r)/s\ -\ l\ /(q \cdot s\ )] = N \cdot \Phi[x/s\ -\ l\ /(q \cdot s\ )] \end{aligned} \qquad (3)$$

where $\Phi(.)$ is the standard normal CDF.

Given that $q_e = f(q_a, \mathbf{R})$, satisfaction of condition WF2 depends on the quality assessment function *f*. More specifically, *f* must be chosen so that, for all steady-state strategies (q_r, q_a) the equation:

$$q_e = q_r + x = f(q_a, \mathbf{R}(x)) \qquad (4)$$

has a unique solution at $x = 0$.

eBay does not specify, or even recommend, a specific quality assessment function *f*. It simply publishes the quantities Σ_+ and Σ_- for each seller and allows buyers to use any assessment rule they see fit. It is important to note at this point that eBay does not currently publish the quantity $\Sigma_{no\ rating}$ (and therefore N) for a seller. As we will show below, knowledge of *N* is essential for constructing reliable quality assessment functions. The results of this paper, therefore, make a strong argument that the number of transactions that have received no rating should be added to the profile information published by eBay and similar systems.

[4] Section 4.2 will explore the conditions under which sellers will indeed find it optimal to settle down to a steady state strategy.

In this paper, we will explore one general family of quality assessment functions which, when implemented correctly and used in conjunction with rating rule (2), satisfy WF2. We will further explore different ways in which users of a reputation mechanism can use Σ_+, Σ_- and N in order to correctly implement those functions.

The general form of our quality assessment functions is given by:

$$q_e = f(q_a, \mathbf{R}) = \begin{cases} q_a & \text{if } \hat{x}(\mathbf{R}) \le 0 \\ 0 & \text{if } \hat{x}(\mathbf{R}) > 0 \end{cases} \qquad (5)$$

In other words, buyers assess the quality of an item to be equal to that advertised by the seller, if, based on the seller's profile, they conclude that the seller *does not over-advertise*. Otherwise, buyers assume that the seller lies and assess minimum quality. Function (5) therefore uses the information provided by the reputation mechanism in order to derive a (binary) assessment of *truthfulness in advertising*.

Assuming that sellers have a way of reliably inferring the sign of x from feedback profile information, sellers who over-advertise their quality will quickly see their estimated quality fall to zero. Therefore, if *f* is given by (5), equation (4) has no solution for $x > 0$.

Note that function (5) does not prevent sellers from under-advertising their quality because for $q_a = q_r + x$, all $x \le 0$ are also solutions of equation (4). However, given that we have assumed that $\partial p / \partial q_a \ge 0$, we would not expect any profit-maximizing seller to under-advertise. Therefore, the only steady-state seller strategy for sellers would be to truthfully advertise their real quality. In that case, buyers would estimate $q_e = q_a = q_r$, a desirable outcome, which satisfies WF2.

Let us now explore three different ways in which buyers can use Σ_+, Σ_- and N in order to estimate the sign of x.

Assessment based on the number of positives

One way to estimate seller honesty is to require that the fraction of positive ratings of good sellers exceed a threshold. From (3) we can see that $\hat{h} \equiv \Sigma_+ / N$ can be interpreted as a point estimator of $\Phi[-x/s]$. Given that $\Phi[-x/s] < 0.5$ for all $x > 0$, assessment of the sign of x reduces to testing the statistical hypothesis $H_0 : h \ge 0.5$ given $\hat{h}$. The corresponding quality assessment function then becomes:

$$q_e = \begin{cases} q_a & \text{if } H_0 \text{ accepted} \\ 0 & \text{if } H_0 \text{ rejected} \end{cases} \qquad (6)$$
$$\text{where } H_0 : h \ge 0.5 \text{ given } \hat{h} \equiv \Sigma_+/\text{N}$$

Hypothesis H_0 can be tested using one of the known techniques for computing confidence intervals of proportions following binomial distributions (e.g. Blyth and Still 1983).

Function (6) is an appealing method for assessing seller quality because of its relative simplicity. Note that its computation does not require knowledge of the model parameters l, q and s . However, (6) is difficult to compute reliably without knowledge of *N*, the total number of rated plus unrated transactions of a seller. As was mentioned, eBay does not make *N* known to its members. Taking $\hat{h} \equiv \Sigma_+ / (\Sigma_+ + \Sigma_-)$ would result in large overestimation of $\Phi[-x / s]$, especially because, due to rating leniency, $\Sigma_{\text{no rating}}$ is expected to be quite significant (in the data set of Figure 1 $\Sigma_{\text{no rating}} / N = 48.3\%$). For that reason, one would infer that quality assessment based on the number of positive ratings is not (and *should not* be) widely used on eBay. This hypothesis is consistent with empirical observations (Dewan and Hsu 2001).

Section 4.2 will discuss another disadvantage of function (6), which is that it makes it easier for sellers to oscillate between periods where they milk their good reputation by overstating their quality and deceiving buyers and periods where they restore their reputation by offering better quality than what buyers expect.

Assessment based on the number of negatives

In an analogous manner, we expect good sellers to have few negative ratings. Therefore, another way to estimate seller honesty is to require that the fraction of negative ratings of good sellers stay below a threshold. From (3) we can see that $\hat{z} \equiv \Sigma_- / N$ can be interpreted as a point estimator of $\Phi[x / s - l / (q \cdot s)]$. Given that $\Phi[x / s - l / (q \cdot s)] > \Phi[-l / (q \cdot s)]$ for all $x > 0$, assessment of the sign of x reduces to testing the statistical hypothesis $H_0' : z \le \Phi[-l / (q \cdot s)]$ given $\hat{z}$. The corresponding quality assessment function then becomes:

$$q_e = \begin{cases} q_a & \text{if } H_0' \text{ accepted} \\ 0 & \text{if } H_0' \text{ rejected} \end{cases} \tag{7}$$
$$\text{where } H_0' : z \le k^* \equiv \Phi[-l / (q \cdot s)] \text{ given } \hat{z} \equiv \Sigma_- / N$$

Let us call $k^* \equiv \Phi[-l / (q \cdot s)]$ the optimum trustworthiness threshold. k^* is a monotonically decreasing function of the leniency factor l . Therefore, the more lenient buyers are when they rate, the lower the threshold of negative ratings to transactions above which they should not trust sellers, and vice versa. This is a result that corresponds well to documented empirical findings: most eBay buyers weigh negative ratings much more heavily than positive ratings when assessing the trustworthiness of a prospective seller (Dewan and Hsu 2001). Given that they seem to be rather lenient when they rate those sellers, according to (7), we would expect them to be strict when assessing the quality of sellers, and therefore to be relatively intolerant of negative ratings. The big question, however, is whether buyers use the right threshold when they judge sellers (in other words, are buyers capable of making the “right” judgment of *just how many negative ratings are too many*?).

From equation (7) we can also see that an optimum k^* can be derived for every l . One way of interpreting this result is that satisfaction of WF2 is always possible no matter how lenient (or strict) buyers are when they rate, *provided that they strike the right balance between rating leniency and quality assessment strictness*. In the next section, we shall prove that, more lenient rating (and correspondingly strict assessment) increases the likelihood that sellers will find it optimal to settle down to a steady-state behavior. Some degree of leniency, therefore, can be beneficial to the stability of the marketplace.

It is also important to point out that, unless buyers use the right threshold k^* when evaluating the number of negative ratings of a seller, WF2 will not be satisfied. If buyers use a threshold $k > k^*$ then there will be some $x > 0$ for which H'_0 will be satisfied and sellers will be able to consistently deceive buyers by over-advertising their quality. In contrast, if $k < k^*$, there will be $x_0 < 0$ such that H'_0 will be rejected for all $x > x_0$. In the latter case, to prevent their estimated quality from dropping to zero, sellers will be forced to under-advertise and, therefore, be consistently under-appreciated by $|x_0|$.

We see, therefore, that the choice of the “right” k^* is crucial to the well functioning of the reputation mechanism, and of the marketplace in general. It is important to ask whether buyers can be reasonably expected to be able to correctly derive it. From equation (7), calculation of k^* requires knowledge of the model parameters l, q and s . It is unlikely that buyers would have accurate understanding and knowledge of those parameters (especially s , which partly reflects properties of the seller). Nevertheless, even if the model parameters are not known, it is possible to estimate the value of $\Phi[-l/(q \cdot s)]$ from Σ_+, Σ_- and N . From (3):

$$\left.\begin{aligned} -x/s &= \Phi^{-1}(\Sigma_+/N) \\ -l/(q \cdot s) + x/s &= \Phi^{-1}(\Sigma_-/N) \end{aligned}\right\} \Rightarrow \quad k^* = \Phi[-l/(q \cdot s)] = \Phi[\Phi^{-1}(\Sigma_+/N) + \Phi^{-1}(\Sigma_-/N)] \qquad (8)$$

If *N* is small then a confidence interval should be constructed for k^*.

Even with the help of equation (8), buyers still need to know *N* in order to properly compute function (7). Overall, function (7) defines a rather fragile rule for assessing seller quality efficiently. Given that a lot of eBay buyers are heavily basing their seller quality assessments on the number of negative ratings on the sellers’ feedback profile, it is very interesting to ask what methods they use to compute their trustworthiness thresholds and, even more important, whether their trustworthiness thresholds do indeed come close to satisfying WF2. Clearly, these are important questions, which invite further empirical and experimental results to complement the results of this work.

N.

4.2 Existence of steady-state behavior

The analysis of Section 4.1 has been based on the assumption that sellers settle down to steady-state real and advertised quality levels. This section will investigate the conditions under which sellers will indeed find it optimal to do so. The alternative is to oscillate between building a good reputation and then milking it by over-advertising real quality. As we argued in Section 3, reputation-mediated marketplaces should be designed in order to induce sellers to settle down to steady state behavior (otherwise information about past behavior will not be very helpful as a way of predicting the future).

The principal result of this section is that when quality assessment is based on functions (7) or (10), which involve negative ratings, then, if the rating leniency factor l is large enough, sellers will find it optimal to settle down to steady state behavior. In contrast, there is no such guarantee when quality assessment is based on function (6), which only involves positive ratings. This result shows that more lenient rating (coupled with more strict quality assessment) supports stability in the system. For the same reason, although more fragile and difficult to "get right", functions (7) and (10), i.e. functions which base seller quality assessment on the number of negative ratings, are preferred to function (6), which only looks at the seller's positive ratings.

In order to derive our result, let us consider ways in which sellers may attempt to realize additional profits through oscillating behavior. Assume that a seller is able to perform N_1 transactions before ratings of those transactions are posted to her feedback profile. This number depends on the frequency of transactions and the delay between transactions and the posting of ratings by buyers (on eBay, this delay is typically 2-3 weeks).

Let us consider a seller who, at the end of period 0, has completed N transactions in the current time window and has accumulated a good reputation, by producing and advertising items of quality q^*, the quality that optimizes profits assuming steady-state behavior. Let us further assume that buyers assess quality based on function (7). At the end of period 0:

$$\left.\frac{\Sigma_-}{N}\right|_{period\ 0} = \frac{\Sigma_-}{N}(x=0) = \Phi[-l/(q\cdot s)] = k^* \qquad (11)$$

At the beginning of period 1 the seller decides to milk her reputation by choosing a real quality q' and then over-advertising her quality by x_1 so that her profit is maximized relative to the steady state case. Given the seller's good past reputation, initially buyers will be deceived. However, after they purchase the seller's items, they will realize their inferior quality and will post proportionally more negative ratings. Therefore, at the end of period 1 (after N_1 "deceiving" transactions):

$$\left.\frac{\Sigma_-}{N}\right|_{period\ 1} = \frac{N\cdot\Phi[-l/(q\cdot s)] + N_1\cdot\Phi[x_1/s - l/(q\cdot s)]}{N+N_1} > k^* \qquad (12)$$

and the seller's subsequent estimated quality will fall to zero. Assuming that some buyers are willing to buy from somebody with zero quality if the price is low enough, our seller will stay in business. In order to increase her reputation once again, she needs to reduce the ratio Σ_-/N to below the threshold k^*. The only way she can achieve this is to go through a period where she produces higher quality items but receives lower prices, surpassing buyers' expectations (who now expect $q_e = 0$) by x_2. Let us assume that it would take N_2 "redeeming" transactions before $\Sigma_-/N \le k^*$. At the end of period 2:

$$\left.\frac{\Sigma_-}{N}\right|_{period\,2} = \frac{N\cdot\Phi[-l/(q\cdot s)] + N_1\cdot\Phi[x_1/s - l/(q\cdot s)] + N_2\cdot\Phi[-x_2/s - l/(q\cdot s)]}{N+N_1+N_2} = k^* = \Phi[-l/(q\cdot s)] \qquad (13)$$

A profit-maximizing seller will choose to oscillate if the profit from the "deceiving" transactions relative to the steady-state profit exceeds the loss from the "redeeming" transactions relative to the steady-state profit. If these two quantities have a finite ratio, then, provided that the number N_2 of "redeeming" transactions that are necessary in order to "undo" the reputation effects of N_1 "deceiving" transactions is high enough, sellers will not find it profitable to oscillate and will settle down to steady-state real and advertised quality levels.

From (13) after some algebraic manipulation, we get:

$$\frac{N_2}{N_1} = \frac{\Phi[x_1 / s - l /(q \cdot s)] - \Phi[-l /(q \cdot s)]}{\Phi[-l /(q \cdot s)] - \Phi[-x_2 / s - l /(q \cdot s)]} = g(l, x_1, x_2) \qquad (14)$$

After some manipulation we get $\partial g / \partial l > 0$ and $\partial^2 g / \partial l^2 > 0$. In fact g(.) grows exponentially with l [5]. Figure 2 plots $g(l)$ for $q = s = 1$ and some representative values of $x = x_1 = x_2$.

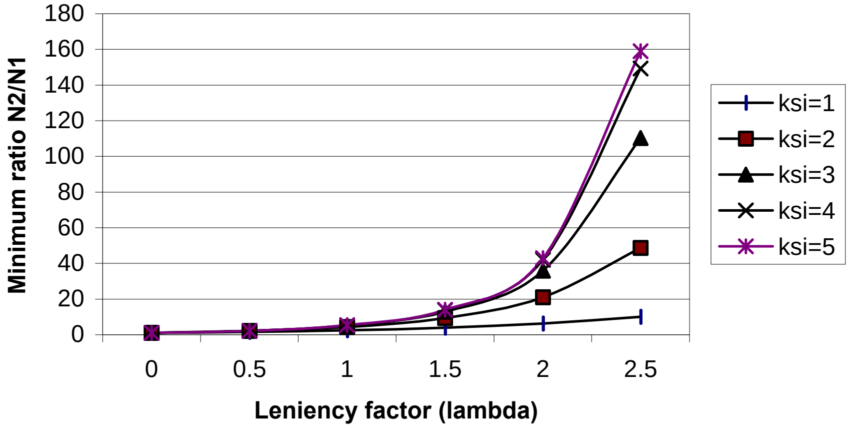

Figure 2: Minimum ratio of "redeeming" to "deceiving" transactions needed in order to restore one's good reputation following a period of quality over-reporting.

From Figure 2 it is evident that in marketplaces where buyers rate leniently (and assess quality strictly), sellers need many more "redeeming transactions" in order to restore their good reputation following a few "deceiving transactions". The relative number of redeeming transactions increases exponentially with the leniency factor. Otherwise said, the larger the l, the more difficult it is for sellers to restore their reputation once they lose it. Consequently, if l is sufficiently large, sellers will find it optimal to settle down to steady-state real and advertised quality levels. Q.E.D.

A similar result can be derived if buyers base quality assessment on the ratio Σ_- / Σ_+. In contrast, if buyers base quality assessment on function (7), our analysis gives:

$$\frac{N_2}{N_1} = \frac{1 - 2 \cdot \Phi[x_1 / s]}{1 - 2 \cdot \Phi[x_2 / s]} \qquad (15)$$

Equation (15) gives $N_2 = N_1$ for $x_1 = x_2$ and N_2 slightly less than N_1 for $x_1 < x_2$. Otherwise said, following a set of deceiving transactions, it takes the same number of (or fewer) redeeming transactions in order to restore one's good reputation. In such a setting, it is

[5] Furthermore, for a given l, g(.) grows rapidly with x_1 and decreases very slowly with x_2. This means that the minimum necessary ratio of redeeming to deceiving transactions grows with the amount of initial deception (x_1) and cannot be significantly brought down by increasing the amount of redemption (x_2).

more likely that some sellers will have profit functions for which it will be optimal to oscillate. Therefore, one expects that in reputation-mediated marketplaces where buyers use (6) to assess seller quality, there will be less stability than in marketplaces where sellers use (7) or (10).

The results of this section provide some interesting arguments for both rating leniency as well as for basing the quality assessment of sellers on their negative, rather than their positive ratings.

5. Reality checks and some recommendations

The results of the previous sections have been derived by making a number of simplifying assumptions about buyer behavior. More specifically, we have assumed that all buyers have the same quality sensitivity q and leniency factor l . Furthermore, we have assumed that buyers *always* submit ratings whenever their satisfaction rises above zero or falls below $-l$. Both assumptions are not likely to hold in a real marketplace. Buyers have different personalities, and therefore, are expected to have different quality sensitivities, as well as leniency parameters. Furthermore, ratings do incur a cost (time to log on and submit them) and some buyers do not bother rating, even when transactions turn out really good or very bad. In this section, we will inject a bit of reality to our model and will explore how our results change if we take into account the above considerations.

Reality Check #1: Some buyers never rate

We need to modify our rating function $r(S)$ in (2), so that when $\mathrm{S} > 0$, $r(S) = "+"$ with probability b and $\mathrm{r(S) = no\ rating}$ with probability $(1-b)$. Similarly, when $S \le -l$, $r(S) = "-"$ with probability g and $\mathrm{r(S) = no\ rating}$ with probability $(1-g)$. Under this new rating function, the statistical hypothesis in (6) becomes $H_0 : h \ge b \cdot 0.5$, while the hypothesis in (7) becomes $H'_0 : z \le g \cdot \Phi[-l /(q \cdot s)]$. We see that our new assumption introduces two additional parameters to our model. The parameters need to be reliably estimated in order for property WF2 to be satisfied.

Reality Check #2: Buyers differ in quality sensitivity and leniency

Let's define $w \equiv l / q$ and let's call $p(w)$ the probability distribution of w among buyers. Then (3) must be modified as:

$$\begin{aligned} \Sigma_+ &= N \cdot \Pr ob[S > 0] = N \cdot \Phi[-x / s] \\ \Sigma_- &= N \cdot \Pr ob[S \le -l] = N \cdot \int \Phi[(x - w) / s] \cdot p(w) \cdot dw \end{aligned} \qquad (16)$$

If quality assessment is based on the fraction of positive ratings using (5), then reality check #2 does not introduce additional complications. However, if quality assessment is based on the fraction of negative ratings, which, in the presence of lenient ratings is the rule most likely to result in stable seller behavior, then things do become considerably more complicated. More specifically, it is easy to see that the statistical hypothesis to be tested in (6) must become $H_0' : z \leq k^* \equiv \int \Phi[-w / s\,] \cdot p(w) \cdot dw$. In order to calculate the "right" k^*, one needs knowledge of $p(w)$.

Things become even more complicated if we combine reality checks #1 and #2, which would be the situation that most closely corresponds to actual reality. Of course, one can begin to think of ways in which individual buyers might be able to estimate, maybe with some degree of error, the additional model parameters b, g and $\mathrm{p}(w)$ from Σ_+, Σ_- and N .

However, instead of embarking in this direction, at this stage we believe that we are provided enough arguments to make one of the main points of this paper: Binary reputation mechanisms can *in theory* be well functioning under the assumption of simple rating and assessment rules, but only if buyers use the right thresholds when judging seller trustworthiness. Calculating the right threshold from Σ_+, Σ_- alone, the only information currently provided by eBay, is very difficult. Calculating the right threshold from Σ_+, Σ_- and N is possible under the simplifying assumptions of Section 2 but becomes more and more difficult as our models approach reality. In realistic cases, the correct assessment rule depends not only on the feedback profile of a seller but also on properties of the rater population. Given that the efficiency of the marketplace crucially depends on the selection of correct assessment thresholds on the part of the buyer, the most sensible course of research therefore should be to think of additional information that the reputation mechanism can provide to raters, in order to make this calculation easier.

One idea is for the operator of the marketplace, or some trusted third party, to introduce a small number of honest sellers into the market. The behavior of those sellers must be completely under the control of the marketplace operator but their identities should be unknown to buyers (so that buyers are not biased when they rate those sellers). Given that the marketplace knows that for those sellers $x = 0$, it can use their feedback profiles in order to derive estimates of b, g , $\mathrm{p}(w)$ and k^* Those estimates, or suitably chosen derivatives, can then be communicated to the buyers for the purpose of facilitating their seller assessment process.

6. Conclusions

The objective of this paper was to explore to what extent binary reputation mechanisms, such as the one used at eBay, are capable of inducing efficient market outcomes in marketplaces where (a) true quality information is unknown to buyers, (b) advertised quality is completely under the control of the seller and (c) the only information available to buyers is an item's advertised quality plus the seller's feedback profile.

The first contribution of the paper is the definition of a set of conditions for evaluating the well functioning of a reputation mechanism is such settings. We consider a reputation mechanism to be well-functioning if it (a) induces sellers to settle down to a steady-state behavior assuming it is optimal for them to do so under perfect quality information and

(b) at steady-state, seller quality as estimated by buyers *before* transactions take place is equal to their true quality.

The second contribution of the paper is an analysis of whether binary reputation mechanisms can be well-functioning under the assumptions that (a) ratings are based on the difference between buyers' true utility following a transaction and their expectations before the transaction and (b) buyers are relatively lenient when they rate and correspondingly strict when they assess a seller's feedback profile.

Our first conclusion is that if binary feedback profiles are used to decide whether a seller advertises truthfully (in which case buyers assess quality equal to the advertised quality) or not (in which case buyers assess quality equal to the minimum quality), then, *in theory*, binary reputation systems can be well functioning, provided that buyers strike the right balance between rating leniency and quality assessment strictness. Furthermore, assuming that buyers base their judgment on the ratio of negative ratings received by a seller, if buyers are lenient enough when they rate and correspondingly strict when they judge seller profiles, we have shown that sellers will find it optimal to settle down to steady-state quality levels if such an equilibrium exists under perfect information. This is an interesting way in which (a) judging seller trustworthiness based on their negative ratings is preferable to basing it on their positive ratings and (b) some degree of rating leniency helps bring stability to the system.

Our second conclusion is that, unless buyers use the "right" threshold parameters when they judge seller profiles, binary reputation mechanisms will not function well and the resulting market outcome will be unfair for either the buyers or the sellers. In that sense, although binary reputation mechanism can be well functioning in theory, they are expected to be quite fragile in practice. The crucial question therefore becomes whether binary feedback profiles provide sellers (esp. relatively unsophisticated ones) with enough information to derive the "right" seller judgment rules.

We have found that the "right" judgment rule (e.g. what is the *right* number of negative ratings above which a seller should not be trusted?) is difficult to infer correctly from knowledge of the sum of positive and negative ratings alone, which is the only information currently provided by eBay to its members. If knowledge of the sum of unrated transactions is added to feedback profiles, then, under a number of simplifying assumptions, it *is* possible to derive non-obvious but relatively simple "optimal" judgment rules which result in well functioning reputation mechanisms. However, if the simplifying assumptions are dropped, calculation of the right judgment rule from Σ_+, Σ_- and N once again becomes difficult, as it requires knowledge not only of seller ratings but of the rater population as well.

Our findings lead to the recommendation that more information should be provided to assist raters of such marketplaces use feedback profiles in the "right" way. One idea is for the operator of the marketplace, or some trusted third party, to introduce a small number of honest sellers under its control into the market. If one knows, a priori, that a seller is honest, her feedback profile can then be used by the marketplace to estimate and publish a number of additional "market-wide" parameters which, together with Σ_+, Σ_- and N , will help buyers more reliably assess the quality of other sellers in the system.

The theoretical results of this paper raise some intriguing questions related to the efficiency, fairness and stability of eBay-like electronic marketplaces. The author would welcome experimental and empirical evidence that will shed more light into the questions raised and would validate the conclusions drawn from his models.

References

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

Blyth, C.R. and Still H.A. (1983) Binomial Confidence Intervals. *Journal of the American Statistical Association* 78, pp. 108-116.

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.

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

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.

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.

Rogerson, W.P. (1983) Reputation and product quality. *Bell Journal of Economics*, Vol. 14 (2), pp. 508-16.

Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. (2000). Analysis of Recommender Algorithms for E-Commerce. *ACM E-Commerce 2000 Conference*, Minneapolis, MN, Oct. 17-20, 2000.

Schmalensee, R. (1978). Advertising and Product Quality. *Journal of Political Economy,* Vol. 86, pp. 485-503.

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.

Smallwood, D. and Conlisk, J. (1979). Product Quality in Markets Where Consumers Are Imperfectly Informed. *Quarterly Journal of Economics*. Vol. 93, pp. 1-23.

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