# A Computational Model of Trust and Reputation

Lik Mui, Mojdeh Mohtashemi,
*200 Technology Square,*
*Cambridge, MA 02139, USA*
*{lmui, mojdeh}@lcs.mit.edu*

Ari Halberstadt
*9 Whittemore Road,*
*Newton, MA 02458, USA*
*ari@magiccookie.com*

## Abstract

*Despite their many advantages, e-Businesses lag behind brick and mortar businesses in several fundamental respects. This paper concerns one of these: relationships based on trust and reputation. Recent studies on simple reputation systems for e-Businesses such as eBay have pointed to the importance of such rating systems for deterring moral hazard and encouraging trusting interactions. However, despite numerous studies on trust and reputation systems, few have taken studies across disciplines to provide an integrated account of these concepts and their relationships. This paper first surveys existing literatures on trust, reputation and a related concept: reciprocity. Based on sociological and biological understandings of these concepts, a computational model is proposed. This model can be implemented in a real system to consistently calculate agents' trust and reputation scores.*

## 1. Introduction

Trust and reputation underlies every face-to-face trade. A major weakness of electronic markets is the raised level of risk associated with the loss of the notions of trust and reputation. In an on-line setting, trading partners have limited information about each other's reliability or the product quality during the transaction. The analysis by Akerloff in 1970 on the Market for Lemons is also applicable to the electronic market. The main issue pointed out by Akerloff about such markets is the information asymmetry between the buyers and sellers. The buyers know about their own trading behavior and the quality of the products they are selling. On the other hand, the sellers can at best guess at what the buyers know from information gathered about them, such as their trustworthiness and reputation. Trading partners use each others' reputations to reduce this information asymmetry so as to facilitate trusting trading relationships.

Reputation reporting systems have been implemented in e-commerce systems such as eBay, Amazon, *etc.*, and have been credited with these systems' successes (Resnick, *et al*., 2000a). Several research reports have found that seller reputation has significant influences on on-line auction prices, especially for high-valued items (Houser and Wooders, 2000; Dewan and Hsu, 2001). Trust between buyers and sellers can be inferred from the reputation that agents have in the system. How this inference is performed is often hand-waved by those designing and analyzing such systems as Zacharia and Maes (1999), Houser and Wooders (2001). Moreover, many studies do not take into account possibilities of deception and distrust. As shown by Dellarocas (2000), several easy attacks on reputation systems can be staged. These studies also do not examine issues related to the ease of changing one's pseudonym online. As Friedman and Resnick (1998) have pointed out, an easily modified pseudonym system creates the incentive to misbehave without paying reputational consequences.

Besides electronic markets, trust and reputation play important roles in distributed systems in general. For example, a trust model features prominently in Zimmermann's Pretty Good Privacy system (Zimmermann, 1995; Khare and Rifkin, 1997). The reputation system in the anonymous storage system Free Haven is responsible for creating accountability of user and component actions (Dingledine, *et al,* 2001). Trust management in the system Publius allows it to publish materials anonymously such that censorship of and tampering with any publication in the system is rendered very difficult (Waldman, *et al.*, 2000).

Despite the obvious usefulness of trust and reputation, conceptual gaps exist in current models about them. Resnick and Zeckhauser (2000b) have pointed out the so called *Pollyanna* effect in their study of the eBay reputation reporting system. This effect refers to the disproportionately positive feedbacks from users and rare negative feedbacks. They have also pointed out that despite the incentives to free ride (for not providing feedbacks), feedbacks by agents are provided in more than half of the transactions. This

violates the rational alternative of taking advantage of the system without spending the effort to provide feedback. Current trust and reputation models cannot account for these observations.

How is "reputation" related to "trust", "image", "propensity to reciprocate" or other related concepts? Fundamentally, reputation is a social concept. This paper attempts to first understand reputation by comparing and contrasting notions of reputation and trust from various social and scientific disciplines. Secondly, this paper proposes a computational model of trust and reputation based on studies across diverse disciplines to provide an integrated account of these concepts and their relationships.

## 2. Understanding Trust and Reputation

Trust and reputation have become important topics of research in many fields. This section reviews a few of the important studies.

Scientometrics refers to the study of measuring research outputs such as journal impact factors. Reputation as used by this community usually refers to number of cross citations that a given author or journal has accumulated over a period of time (Garfield, 1955). As pointed out by Makino, *et al.*, 1998 and others, cross citation is a reasonable but sometimes confounded measure of one's reputation.

Economists have studied reputation in game theoretic settings. Entry deterrence is one of the early areas for game theorists' study of reputation. Kreps and Wilson (1982) postulate that imperfect information about players' payoffs creates "reputation effects" for multi-stage games. They claim that an incumbent firm seeks to acquire an early reputation for being "tough" in order to decrease the probability for future entries into the industry. Milgrom and Roberts (1982) report similar findings by using asymmetric information to explain the reputation phenomenon. For an incumbent firm, it is rational to seek a "predation" strategy for early entrants even if "it is costly when viewed in isolation, because it yields a reputation which deters other entrants." (*ibid*.)

In the computer science literature, Marsh (1994) is among the first to introduce a computational model for trust in the distributed artificial intelligence (DAI) community. He did not model reputation in his work. As he has pointed out, several limitations exist for his simple trust model. Firstly, trust is represented in his model as a subjective real number between the arbitrary range –1 and +1. The model exhibits problems at the extreme values and at 0. Secondly, the operators and algebra for manipulating trust values are limited and have trouble dealing with negative trust values. Marsh also pointed to difficulties with the concept of "negative" trust and its propagation.

Zacharia and Maes (1999) have suggested that reputation in an on-line community can be related to the ratings that an agent receives from others, and have pointed out several criteria for such rating systems. Their mathematical formulation for the calculation of reputation can at best be described as intuitive – without justifications except the intuitive appeal of the resulting reputation dynamics.

Abdul-Rahman, *et al,* (2000) have proposed that the trust concept can be divided into *direct* and *recommender* trust. They represent direct trust as one of four agent-specified values about another agent ("very trustworthy", "trustworthy", "untrustworthy", and "very untrustworthy"). Recommended trust can be derived from word-of-mouth recommendations, which they consider as "reputation". The translation from recommendations to trust is performed through an *ad-hoc* scheme. *Ad-hoc* formulation plagues several other proposals for reputation/trust systems such as those in Glass, *et al.* (2000), Yu and Singh (2001), Esfandiari, *et al.*, (2001), Rouchier, *et al.* (2001), Sabater, *et al.*, (2001), among others. Nevertheless, reputation and trust have been found to provide useful intuition or services for of these systems.

Whether online reputation systems contribute to trade is answered by several research analysis of existing systems. Resnick and Zeckhauser (2000b) have analyzed the feedback rating system used in eBay as a reputation system. "Reputation" is taken to be a function of the cumulative positive and non-positive ratings for a seller or buyer. Trust by one agent of another is inferred by an implicit mechanism. They have found that the system does encourage transactions.

Houser and Wooders (2000) have studied auctions in eBay and describe reputation as the *propensities to default* – for a buyer, it is the probability that if the buyer wins, he will deliver the payment as promised before the close of the auction; for a seller, it is the probability that once payment is received, he will deliver the item auctioned. Their economic analysis shows that reputation has a statistically significant effect on price. Unfortunately, they did not model how reputation is built; nor how trust is derived from reputation.

Both Lucking-Reily, *et al.* (1999) and Bajari and Hortacsu (2000) have examined coin auctions in eBay. These economic studies have provided empirical confirmation of reputation effects in internet auctions. Bajari and Hortacsu (2000) have also reported the "winner's curse" phenomenon in their analysis. This phenomenon refers to a fall in the bidder's expected profits when the expected number of bidders is increased.

"Be nice to others who are nice to you" seems to be a social dictum well permeated in our society for encouraging social cooperation. It is also very much related to trust and reputation, as well as to the concept of "reciprocity" as studied by evolutionary biologists. Trivers (1971) has suggested the idea of *reciprocal altruism* as an explanation for the evolution of cooperation. Altruists indirectly contribute to their fitness (for reproduction) through others who reciprocate back. Reputation and trust can potentially help to distinguish altruists from those disguised as such, thereby preventing those in disguise from exploiting the altruists. Alexander (1987) greatly extended this idea to the notion of *indirect reciprocity*. In situations involving cooperators and defectors, *indirect reciprocity* refers to reciprocating toward cooperators indirectly through a third party. Indirect reciprocity "…involves reputation and status, and results in everyone in the group continually being assessed and reassessed." Alexander has argued that indirect reciprocity (and reputation and status) is integral to the proper functioning of human societies.

Nowak and Sigmund (1998, 2000) use the term *image* to denote the total points gained by a player by reciprocation. The implication is that image is equal to reputation. Image score is accumulated (or decremented) for direct interaction among agents. Following the studies by Pollock and Dugatkin (1992), Nowak and Sigmund (1998) have also studied the effects of observers on image scores. Observers have a positive effect on the development of cooperation by facilitating the propagation of observed behavior (image) across a population. Castelfranchi, *et al.* (1998) explicitly have reported that communication about "Cheaters"'s bad reputation in a simulated society is vital to the fitness of agents who prefer to cooperate with others.

Among sociologists, reputation as a quantitative concept is often studied as a network parameter associated with a society of agents (Wasserman and Faust, 1994). Reputation or prestige is often measured by various centrality measures. An example is a measure proposed by Katz (1953) based on a stochastic coincidence matrix where entries record social linkages among agents. Because the matrix is stochastic, the right eigenvector associated with the eigenvalue of 1 is the stationary distribution associated with the stochastic matrix (Strang, 1988). The values in the eigenvector represent the reputations of the individuals in the society. Unfortunately, these values are often global in nature, and lacks context dependence.

In summary, the trust and reputation studies examined so far have exhibited one or more of the following weaknesses:

- Differentiation of trust and reputation is either not made or the mechanism for inference between them is not explicit.
- Trust and reputation are taken to be the same across multiple contexts or are treated as uniform across time.
- Despite the strong sociological foundation for the concepts of trust and reputation, existing computational models for them are often not grounded on understood social characteristics of these quantities.

This paper proposes a computational model that attempts to address the concerns raised here.

## 3. Model Rationale

Contrary to game theorists' assumptions that individuals are rational economic agents [1] who use backward induction to maximize private utilities (Fudenberg and Tirole, 1996; Binmore, 1997), field studies show that individuals are boundedly rational [2] (Simon, 1996) and do not use backward induction in selecting actions [3] (Rapoport, 1997; Hardin, 1997). Social-biologists and psychologists have shown in field studies that humans can effectively learn and use heuristics [4] in decision making (Barkow, *et al.*, 1992; Guth and Kliemt, 1996; Trivers, 1971). One important heuristics that has been found to pervade human societies is *reciprocity norm* for repeated interactions with the same parties (Becker, 1990; Gouldner, 1960). In fact, people use reciprocity norms even in very short time-horizon interactions (McCabe, *et al*., 1996). *Reciprocity norms* refer to social strategies that individuals learn which prompt them to "… react to the positive actions of others with positives responses and the negative actions of others with negative responses (Ostrom, 1998). From common day experience, we know that the degree to which reciprocity is expected and used is highly variable from one individual to another. Learning the degree to which reciprocity is expected can be posed as a trust estimation problem.

---

[1] Rational agents refer to those able to deliberate, *ad infinitum*, the best choice (for maximizing their private utility functions) without regard to computational limitations (*c.f.*, Fudenberg and Tirole, 1991).

[2] Bounded rationality refers to rationality up to limited computational capabilities (*c.f.* Simon, 1981)

[3] Backward induction here refers to a style of inference based on inducting from the last game of a sequence of games by maximizing a given utility at each step (this style can also be characterized as dynamic programming) (*c.f.*, Axelrod, 1984; Fudenberg and Tirole, 1996).

[4] A heuristic refers to "rules of thumb — that [individuals] have learned over time regarding responses that tend to give them good outcomes in particular kinds of situations." (Ostrom, 1998)

There are many reciprocity strategies proposed by game-theoreticians; the most famous of which is the tit-for-tat strategy which has been extensively studied in the context of the Prisoners's Dilemma game (Axelrod, 1984; Pollock and Dugatkin, 1992; Nowak and Sigmund, 2000). Not everyone in a society learns the same norms in all situations. Structural variables affect individuals' level of confidence and willingness to reciprocate. In the case of cooperation, some cooperate only in contexts where they expect reciprocation from their interacting parties. Others will only do so when they are publicly committed to an agreement.

When facing social dilemmas [5], trustworthy individuals tend to **trust** others with a reputation for being trustworthy and shun those deemed less so (Cosmides and Tooby, 1992). In an environment where individuals "regularly" perform **reciprocity** norms, there is an incentive to acquire a **reputation** for reciprocative actions (Kreps, 1990; Milgrom, *et al.*, 1990; Ostrom, 1998). "Regularly" refers to a *caveat* observed by sociologists that reputation only serves a normative function in improving the fitness of those who cooperate while disciplining those who defect if the environment encourages the spreading of reputation information (Castelfranchi, *et al.*, 1998). In the words of evolutionary biologists, having a good reputation increases an agent's *fitness* in an environment where reciprocity norms are expected (Nowak and Sigmund, 1998). Therefore, developing the quality for being trustworthy is an asset since trust affects how willing individuals are to participate in reciprocative interactions (Dasgupta, 2000; Tadelis, 1999).

The following section will transform these statements into mathematical expressions. The intuition behind the model given here is inspired by Ostrom's 1998 Presidential Speech to the American Political Society, which proposed a qualitative behavioral model for collective action.

To facilitate the model description, agents and their environment are to be defined. Consider the scenario that agent $a_j$ is evaluating $a_i$'s reputation for being cooperative. The set of all agents that $a_j$ asks for this evaluation can be considered to be a unique society of $N$ agents **A** (where both the elements in **A** and its size depend on different $a_j$'s). **A** is called an "embedded social network" with respect to $a_j$ (Granovetter, 1985):

***Agents***: $\mathbf{A} = \{a_1, a_2, \ldots a_N\}$

[5] Social dilemma refers to a class of sociological situations where maximization of personal utilities do not necessarily lead to the most desirable outcome. Tragedy of the commons (Hardin, 1968) or Prisoner's dilemma (Axelrod, 1984) is the most studied social dilemma.

The reputation of an agent $a_i$ is *relative* to the particular embedded social network in which $a_i$ is being evaluated.

It should be clear from the argument thus far that reciprocity, trust and reputation are highly related concepts. The following relationships are expected:

- Increase in agent $a_i$'s reputation in its embedded social network **A** should also increase the trust from the other agents for $a_i$.
- Increase in an agent $a_j$'s trust of $a_i$ should also increase the likelihood that $a_j$ will reciprocate positively to $a_i$'s action.
- Increase in $a_i$'s reciprocating actions to other agents in its embedded social network **A** should also increase $a_i$'s reputation in **A**.

Decrease in any of the three variables should lead to the reverse effects. Graphically, these intuitive statements create the following relationships among the three variables of interest:

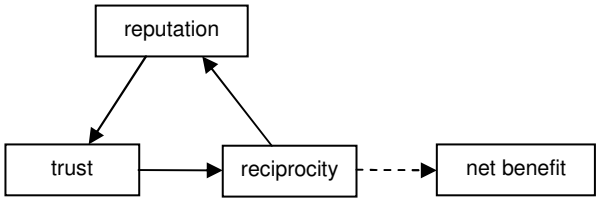


*Figure 1. This simple model shows the reinforcing relationships among trust, reputation and reciprocity. The direction of the arrow indicates the direction of influence among the variables. The dashed line indicates a mechanism not discussed.*[6]

This paper uses the following definition for reciprocity:

- ***Reciprocity***: mutual exchange of deeds (such as favor or revenge).

This definition is largely motivated by the many studies of reciprocity in which repeated games are played between two or more individuals (Raub and Weesie, 1990; Boyd and Richersen 1989; Nowak and Sigmund, 1998). Two types of reciprocity are considered: direct reciprocity refers to interchange between two concerned agents; indirect reciprocity refers to interchange between two concerned agents interceded by mediating agents in between.

Reciprocity can be measured in two ways. Firstly, reciprocity can be viewed as a social norm shared by agents in a society. The higher this "societal reciprocity," the more likely one expects a randomly selected agent from that society to engage in

[6] Ostrom (1998) discusses how reciprocity affects the level of cooperation which affects the overall net benefits in a society.

reciprocating actions. Secondly, reciprocity can be viewed as a dyadic variable between two agents (say $a_i$ and $a_j$). The higher this "dyadic reciprocity," the more one expects $a_i$ and $a_j$ to reciprocate each other's actions. In this latter case, no expectation about other agents should be conveyed. For any single agent $a_i$, the cumulative dyadic reciprocity that $a_i$ engages in with other agents in a society should have an influence on $a_i$'s *reputation* as a reciprocating agent in that society.

- ***Reputation***: perception that an agent creates through past actions about its intentions and norms [7]

Reputation is a social quantity calculated based on actions by a given agent $a_i$ and observations made by others in an "embedded social network" that $a_i$ resides (Granovetter, 1985). $a_i$'s reputation clearly affects the amount of trust that others have toward it. How is trust defined?

The definition for trust by Gambetta (1988) is often quoted in the literature: "… trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent will perform a particular action, both before [it] can monitor such action (or independently of his capacity of ever to be able to monitor it) and in a context in which it affects [the agent's] own action" (*ibid.*). This paper elects the term "subjective expectation" rather than "subjective probability" to emphasize the point that trust is a summary quantity that an agent has toward another based on *a number of former encounters* between them:

- ***Trust***: a subjective expectation an agent has about another's future behavior based on the history of their encounters.

Trust is a subjective quantity calculated based on the two agents concerned in a dyadic encounter. Dasgupta (2000) gave a similar definition for trust: the expectation of one person about the actions of others that affects the first person's choice, when an action must be taken before the actions of others are known.

Given the simple model of interaction in *Figure 1*, the rest of this paper operationalizes this model into mathematical statements that can be implemented in a real world system.

## 4. Notations for Model

To simplify the reasoning about the main quantities of interest (reciprocity, trust, and reputation), two simplifications are made in this paper. First, the embedded social networks in which agents are embedded are taken to be static. *i.e.*, no new agents are expected to join or leave. Secondly, the action space is restricted to be:

***Action***: $\alpha \in \{ \textit{cooperate}, \textit{defect} \}$

In other words, only binary actions are considered. Let $0 < \gamma < 1$ represents the level of reciprocity norm in the embedded social network where low $\gamma$ represents low level of reciprocity and *vice versa*:

***Reciprocity***: $\gamma \in [0, 1]$

$\gamma$ measures the amount of reciprocative actions that occur in a society. In other words, "*cooperate*" actions are met with "*cooperate*" response; "*defect*" actions are met with "*defect*" responses. How $\gamma$ is derived in our model will be discussed shortly.

Let **C** be the set of all contexts of interest. The reputation of an agent is a social quantity that varies with time. Let $\theta_{ji}(c)$ represent $a_i$'s reputation in an embedded social network of concern to $a_j$ for the context $c \in \mathbf{C}$. In this sense, reputation for $a_i$ is subjective to every other agent since the embedded social network that connects $a_i$ and $a_j$ is different for every different $a_j$. Reputation is the perception that suggests an agent's intentions and norms in the embedded social network that connects $a_i$ and $a_j$. $\theta_{ji}(c)$ measures the likelihood that $a_i$ reciprocates $a_j$'s actions, and can be reasonably represented by a probability measure:

***Reputation***: $\theta_{ji}(c) \in [0, 1]$

Low $\theta_{ji}(c)$ values confer low intention to reciprocate and high values indicate otherwise. As agent $a_i$ interacts with $a_j$, the quantity $\theta_{ji}(c)$ as estimated by $a_j$ is updated with time as $a_j$'s perception about $a_i$ changes.

To model interactions among agents, the concept of an encounter between two agents is necessary. An encounter is an event between two agents ($a_i$, $a_j$) within a specific context such that $a_i$ performs action $\alpha_i$ and $a_j$ performs action $\alpha_j$. Let **E** represent the set of encounters. This set is characterized by:

***Encounter***: $e \in \mathbf{E} = \alpha^2 \blacktriangleright \mathbf{C} \blacktriangleright \{ \perp \}$

where $\{\perp\}$ represents the set of no encounter ("bottom"). While evaluating the trustworthiness of $a_i$, any evaluating agent $a_j$ relies on its knowledge about $a_i$ garnered from former encounters or hearsay about $a_i$. Let $\mathrm{D}_{ji}(c)$ represents a history of encounters that $a_j$ has with $a_i$ within the context $c$:

***History***: $\mathrm{D}_{ji}(c) = \{\mathbf{E}^*\}$

where * represents the Kleene closure, and $D_{ji}$ might include observed encounters involving other agents' encounters with $a_i$. Based on $\mathrm{D}_{ji}(c)$, $a_j$ can calculate its

[7] Ostrom (1998) defines norm as "… heuristics that individuals adopt from a moral perspective, in that these are the kinds of actions they wish to follow in living their life."

trust toward $a_i$, which expresses $a_j$'s expectation of $a_i$'s intention for reciprocation. The above statement can be translated to a pseudo-mathematical expression (which is explained latter in the paper):

***Trust***: $\tau(c) = E[\,\theta(c) \mid D(c)\,]$

The higher the trust level for agent $a_i$, the higher the expectation that $a_i$ will reciprocate agent $a_j$'s actions.

## 5. Computational Model

Consider two agents $a$ and $b$, and assume that they care about each others' actions within a specific context $c$. For clarity, a single context '$c$' is used for all variables. To be estimated is $b$'s reputation in the eyes of $a$: $\theta_{ab}$. In this discussion, we take the viewpoint that $a$ always perform "cooperate" actions and that $a$ is assessing $b$'s tendency to reciprocate cooperative actions. Let a binary random variable $x_{ab}(i)$ represent the $i$th encounter between $a$ and $b$. $x_{ab}(i)$ takes on the value '1' if $b$'s action is 'cooperate' (with $a$) and '0' otherwise. Let the set of $n$ previous encounters between $a$ and $b$ be represented by:[8]

***History***: $D_{ab} = \{\, x_{ab}(1), x_{ab}(2), \ldots, x_{ab}(n) \,\}$

Let $p$ be the number of cooperation by agent $b$ toward $a$ in the $n$ previous encounters. $b$'s reputation $\theta_{ab}$ for agent $\underline{a}$ should be a function of both $p$ and $n$. A simple function can be the proportion of cooperative action over all $n$ encounters. From statistics, a *proportion* random variable can be modeled as a Beta distribution (Dudewicz and Mishra, 1988): $p(\hat{\theta}) = Beta(c_1, c_2)$ where $\hat{\theta}$ represents an estimator for $\theta$, and $c_1$ and $c_2$ are parameters determined by prior assumptions — as discussed later in this section. This proportion of cooperation in $n$ finite encounters becomes a simple estimator for $\theta_{ab}$:

$$\hat{\theta}_{ab} = \frac{p}{n}$$

Assuming that each encounter's cooperation probability is independent of other encounters between $a$ and $b$, the likelihood of $p$ cooperations and $(n - p)$ defections can be modeled as: $L(D_{ab} \mid \hat{\theta}) = \hat{\theta}^{p}(1-\hat{\theta})^{n-p}$. The Beta distribution turns out to be the conjugate prior for this likelihood (Heckerman, 1996). Combining the prior and the likelihood, the posterior estimate for $\hat{\theta}$ becomes (the subscripts are omitted):

$$p(\hat{\theta} \mid D) = Beta(c_1 + p, c_2 + n - p)$$

The steps of derivation for this formula are given in (*Mui, et al. 2001*). First order statistical properties of the posterior are summarized below for the posterior estimate of $\hat{\theta}$:

$$E[\hat{\theta} \mid D] = \frac{c_1 + p}{c_1 + c_2 + n} \qquad \sigma_{\hat{\theta}|D}^{\,2} = \frac{(c_1 + p)(c_2 + n - p)}{(c_1 + c_2 + n - 1)(c_1 + c_2 + n)^2}$$

In their next encounter, $a$'s estimate of the probability that $b$ will cooperate can be shown to be (*ibid.*):

$$\tau_{ab} = p(\, x_{ab}(n+1) = 1 \mid D\,) = E[\,\hat{\theta} \mid D\,]$$

Based on our model shown in *Figure 1*, **trust** toward $b$ from $a$ is this conditional expectation of $\hat{\theta}$ given $D$. The following theorem provides a bound on the parameter estimate $\hat{\theta}$.

***Theorem*** *(Chernoff Bound).* Let $x_{ab}(1), x_{ab}(2), \ldots x_{ab}(m)$ be a sequence of m independent Bernoulli trials,[9] each with probability of success $E(x_{ab}) = \theta$. Define the following estimator:

$$\hat{\theta} = \left(x_{ab}(1) + x_{ab}(2) + \cdots + x_{ab}(m)\right)/m$$

$\hat{\theta}$ is a random variable representing the portion of success, so $E[\hat{\theta}] = \theta$ . Then for $0 \le \varepsilon \le 1$ and $0 \le \delta \le 1$, the following bound hold:

$$Pr\left[\left|\theta - \hat{\theta}\right| \ge \varepsilon\right] \le 2e^{-2m\varepsilon^2} \le \delta \qquad \square$$

The proof is a straightforward application of the additive form of the Chernoff (Hoeffding) Bound for Bernoulli trials (Ross, 1995). Note that "success" in the theorem refers to cooperation in our example, but to reciprocation in general. Also note that $\varepsilon$ refers to the deviation of the estimator from the actual parameter. In this sense, $\varepsilon$ can be considered as a fixed *error* parameter (*e.g.*, 0.05).

From the theorem, $m$ represents the minimum number of encounters necessary to achieve the desired level of confidence and error. This minimum bound can be calculated as follows:

$$m \ge -\frac{1}{2\varepsilon^2}\ln\left(\delta/2\right)$$

Let $\gamma_c = 1 - \delta$. $\gamma_c$ is a confidence measure on the estimate $\hat{\theta}$. A $\gamma_c$ approaches 1, a larger $m$ is required to achieve a given level of error bound $\varepsilon$. $\gamma_c$ can be chosen exogenously to indicate an agent's level of confidence for the estimated parameters.

[8] For clarify, the discussion takes the viewpoint of "direct" encounters between $a$ and $b$. It is equally sensible to include observed encounters about $a$'s actions toward others.

[9] The independent Bernoulli assumption made here for the sequence of encounters is unrealistic for repeated interactions between two agents. Refinements based on removing this assumption are work in progress.

In our model, **reciprocity** represents a measure of reciprocative actions among agents. A sensible measure for "dyadic reciprocity" is the proportion of the total number of cooperation/cooperation and defection/defection actions over all encounters between two agents. Similarly, "societal reciprocity" can be expressed as the proportion of the total number of cooperation/cooperation and defection/defection actions over all encounters in a social network. All encounters are assumed to be dyadic; encounters involving more than two agents are not modeled.

Let $\gamma_{ab}$ represent the measured dyadic reciprocity between agent $a$ and $b$. If $\gamma_{ab} < \gamma_c$, calculated reputation and trust estimates fall below the exogenously determined critical value $\gamma_c$ and are not reliable.

***Complete Stranger Prior Assumption***

If agents $a$ and $b$ are complete strangers — with no previous encounters and no mutually known friends, an ignorance assumption is made. When these two strangers first meet, their estimate for each other's reputation is assumed to be uniformly distributed across the reputation's domain:

$$p(\hat{\theta}) = \begin{cases} 1 & 0 < \hat{\theta} < 1 \\ 0 & otherwise \end{cases}$$

For the Beta prior, values of $c_1$=1 and $c_2$=1 yields such a uniform distribution.

## 6. Propagation Mechanism for Reputation

The last section has considered how reputation can be determined when two agents are concerned. This section extends the analysis to arbitrary number of agents. A schematic diagram of an embedded social network for agents $a$ and $b$ is shown in the figure below: [10]

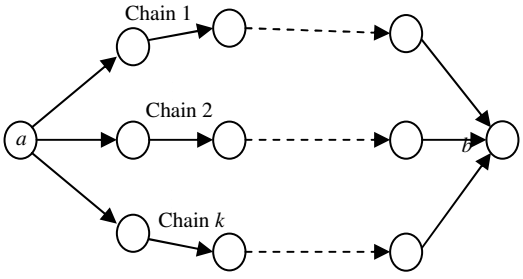


*Figure 2. Illustration of a parallel network between two agents a and b.*

*Figure 2* shows a parallel network of $k$ chains between two agents of interest, where each chain consists of at least one link. Agent $a$ would like to estimate agent $b$'s reputation as defined by the embedded network between them. [11] Clearly, to combine the parallel evidence about $b$, measures of "reliability" are required to weight all the evidences.

From the last section, a threshold ($m$) can be set on the number of encounters between agents such that a reliability measure can be established as follows:

$$w_{ab} = \begin{cases} \dfrac{m_{ab}}{m} & if\ m_{ab} < m \\ 1 & otherwise \end{cases}$$

where $m_{ab}$ is the number of encounters between agents $a$ and $b$. Intuition of this formula is as follows: arguments by Chernoff bound in the last section have established a formula to calculate the minimum sample size of encounters to reach a confidence (and error) level about the estimators. Above a given level of sample size, the estimator is guaranteed to yield the specified level of confidence. Therefore, such an estimate can be considered as "reliable" with respect to the confidence specification. Any sample size less than the threshold $s$ is bound to yield less reliable estimates. As a first order approximation, a linear drop-off in reliability is assumed here.

For each chain in the parallel network, how should the total weight be tallied? Two possible methods are plausible: additive and multiplicative. The problem with additive weight is that if the chain is "broken" by a highly unreliable link, the effect of that unreliability is local to the immediate agents around it. In a long social chain however, an unreliability chain is certain to cast serious doubt on the reliability of any estimate taken from the chain as a whole. On the other hand, a multiplicative weighting has "long-distance" effect in that an unreliable link affects any estimate based on a path crossing that link. The form of a multiplicative estimate for chain $i$'s weight ($w_i$) can be:

$$w_i = \prod_{j=1}^{l_i} w_{ij} \qquad where\ 0 \le i \le k$$

where $l_i$ refers to the total number of edges in chain $i$ and $w_{ij}$ refers to the $j^{th}$ segment of the $i^{th}$ chain.

Once the weights of all chains of the parallel network between the two end nodes are calculated, the estimate across the whole parallel network can be sensibly expressed as a weighted sum across all the chains:

[10] "Embedded social network" refers to the earlier discussion in Section 3.

[11] In general, embedded social networks do not form non-overlapping parallel chains. This arbitrary graph case is discussed in a forthcoming paper.

$$r_{ab} = \sum_{i=1}^{k} r_{ab}(i)\overline{w_i}$$

where $r_{ab}(i)$ is *a*'s estimate of *b*'reputation using path *i* and $\overline{w_i}$ is the normalized weight of path *i* ( $\overline{w_i}$ sum over all *i* yields 1). $r_{ab}$ can be interpreted as the overall perception that *a* garnered about *b* using all paths connecting the two.

## 7. Conclusion

This paper has surveyed the literatures on trust and reputation models across diverse disciplines. A number of significant shortcomings of these models have been pointed out. We have attempted to integrate our understanding across the surveyed literatures to construct a computational model of trust and reputation. Our model has the following characteristics:

- makes explicit the difference between trust and reputation
- defines reputation as a quantity relative to the particular embedded social network of the evaluating agent and encounter history
- defines trust as a dyadic quantity between the trustor and the trustee which can be inferred from reputation data about the trustee
- proposes a probabilistic mechanism for inference among trust, reputation, and level of reciprocity

The explicit formulation of trust, reputation, and related quantities suggests a straightforward implementation of the model in a multi-agent environment (such as an electronic market).

Two immediate future works follow what is presented in this paper. Firstly, the propagation mechanism for reputation only applies to parallel networks. Extending the mechanism to arbitrary graphs with reasonable computational complexity would generalize the model proposed here. A forthcoming paper addresses this mechanism. Secondly, although context is explicitly modeled in the parameters studied here, cross-contexts estimation for the parameters in our model is not addressed. A simple scheme is to create vectorized versions of the quantities studied in this paper. More complex schemes would involve semantic inferences across different contexts.

## 8. References


[1] A. Abdul-Rahman, S. Hailes (2000) "Supporting Trust in Virtual Communities," *Proc. 33rd Hawaii International Conference on System Sciences*.

[2] G. Akerlof (1970) "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism," *Quarterly Journal of Economics, 84*, pp. 488-500.

[3] R. Axelrod (1984) *The Evolution of Cooperation*. New York: Basic Books.

[4] P. Bajari, A. Hortacsu (1999) "Winner's Curse, Reserve Prices and Endogenous entry: Empirical Insights from eBay Auctions," *Stanford Institute for Economic Policy Research (SIEPR) Policy paper No. 99-23*.

[5] J. H. Barkow, L. Cosmides, J. Tooby (eds.) (1992) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press

[6] L. C. Becker, (1990) *Reciprocity*. Chicago: University of Chicago Press.

[7] K. Binmore (1997) "Rationality and Backward Induction," *Journal of Economic Methodology*, *4*, pp. 23-41.

[8] R. Boyd and P. J. Richerson (1989) "The Evolution of Indirect Reciprocity," *Social Networks, 11*, pp. 213-236.

[9] C. Castelfranchi, R. Conte, M. Paolucci (1998) "Normative Reputation and the Costs of Compliance," *Journal of Artificial Societies and Social Simulations, 1*(3).

[10] L. Cosmides, J. Tooby (1992) "Cognitive Adaptations for Social Exchange," in J. H. Barkow, L. Cosmides, J. Tooby (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture,* New York: Oxford University Press, pp. 163-228.

[11] P. Dasgupta (2000) "Trust as a Commodity," in D. Gambetta (ed.) *Trust: Making and Breaking Cooperative Relations, electronic edition*, Department of Sociology, University of Oxford.

[12] S. Dewan, V. Hsu (2001) "Trust in Electronic Markets: Price Discovery in Generalist Versus Specialty Online Auctions," *working paper: http://databases.si.umich.edu /reputations/bib/papers/Dewan&Hsu.doc*.

[13] C. Dellarocas (2000) "Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior," *Proc. 2nd ACM Conference on Electronic Commerce*.

[14] R. Dingledine, M. J. Freedman, D. Molnar (2001) "Free Haven," *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*, O'Reilly.

[15] B. Esfandiari, S. Chandrasekharan (2001) "On How Agents Make Friends: Mechanisms for Trust Acquisition," *4th Workshop on Deception, Fraud and Trust in Agent Societies*, Montreal, Canada.

[16] E. Friedman, P. Resnick (1998) "The Social Cost of Cheap Pseudonyms," *Telecommunications Policy Research Conference*.

[17] D. Fudenberg, J. Tirole (1991) *Game Theory*, Cambridge, Massachusetts: MIT Press.

[18] D. Gambetta (1988) *Trust: Making and Breaking Cooperative Relations*, Oxford: Basil Blackwell.

[19] A. Glass, B. Grosz (2000) "Socially Conscious Decision-Making," *Autonomous Agents'2000*.

[20] E. Garfield (1955) "Citation Indexes for Science," *Science, 122*, pp. 108-111.

[21] A. W. Gouldner (1960) "The Norm of Reciprocity: A Preliminary Statement," *American Sociological Review, 25*, pp. 161-78.

[22] M. Granovetter (1985) “Economic Action and Social Structure: the Problem of Embeddedness,” *American Journal of Sociology, 91*, pp. 481-510.
[23] W. Guth, H. Kliemt (1998) “The Indirect Evolutionary Approach: Bridging the Gap between Rationality and Adaptation,” *Rationality and Society, 10* (3), pp. 377 – 399.
[24] G. Hardin (1968) "The Tragedy of the Commons," *Science 162*(1), pp. 243-48.
[25] R. Hardin (1997) "Economic Theories of the State," in D. C. Mueller (ed.) *Perspectives on Public Choice: A Handbook*, Cambridge: Cambridge University Press, pp. 21-34.
[26] D. Heckerman (1996) “A Tutorial on Learning with Bayesian Networks,” *Technical Report MSR-TR-95-06*, Microsoft Research.
[27] D. E. Houser and J. Wooders (2001) “Reputation in Internet Auctions: Theory and Evidence from eBay,” *working paper: http://w3.arizona.edu/~econ/working_papers /Internet_Auctions.pdf*.
[28] L. Katz (1953) “A New Status Index Derived from Sociometric Analysis,” *Psychometrika, 18*, pp. 39-43.
[29] R. Khare, A. Rifkin (1997) "Weaving a Web of Trust," *World Wide Web Journal, 2*(3), pp. 77-112.
[30] D. M. Kreps, R. Wilson (1982) “Reputation and Imperfect Information,” *Journal of Economic Theory, 27*, pp. 253-279.
[31] D. M. Kreps (1990) "Corporate Culture and Economic Theory," in J. E. Alt, K. A. Shepsle (eds.) *Perspectives on Positive Political Economy*, New York: Cambridge University Press, pp. 90-143.
[32] D. Lucking-Reiley, D. Bryan, N. Prasa, D. Reeves (1999) “Pennies from eBay: The Determinants of Price in Online Auctions,” working paper: http://eller.arizona.edu/~reiley/papers/PenniesFromEBay.pdf
[33] K. A. McCabe, S. J. Rassenti, V. L. Smith (1996) “Game Theory and Reciprocity in Some Extensive Form Experimental Games,” *Proceedings of the National Academy of Sciences*, 9313421-13428
[34] J. Makino, Y. Fujigaki, and Y. Imai (1997) “Productivity of Research Groups – Relation between Citation Analysis and Reputation within Research Community,” *Japan Journal of Science, Technology and Society, 7*, pp. 85-100.
[35] S. Marsh (1994) *Formalising Trust as a Computational Concept,* Ph.D. Thesis, University of Stirling.
[36] P. R. Milgrom, J. Roberts (1982) “Predation, Reputation and Entry Deterrence,” *Journal of Economic Theory, 27*, pp. 280-312.
[37] P. R. Milgrom, D. C. North, 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), pp. 1-23.
[38] L. Mui, M. Mohtashemi, C. Ang, P. Szolovits, A. Halberstadt (2001) “Bayesian Ratings in Distributed Systems: Theories, Models, and Simulations,” *MIT LCS Memorandum*.
[39] M. A. Nowak, and K. Sigmund (1998) “Evolution of Indirect Reciprocity by Image Scoring,” *Nature, 393*, pp. 573-577.
[40] M. A. Nowak, and K. Sigmund (2000) “Cooperation versus Competition,” *Financial Analyst Journal, July/August*, pp. 13-22.
[41] E. Ostrom (1998) “A Behavioral Approach to the Rational-Choice Theory of Collective Action,” *American Political Science Review, 92*(1), pp. 1-22.
[42] G. B. Pollock, L. A. Dugatkin (1992) “Reciprocity and the Evolution of Reputation,” *Journal of Theoretical Biology, 159*, pp. 25-37.
[43] W. Raub, J. Weesie (1990) “Reputation and Efficiency in Social Interactions: An Example of Network Effects,” *American Journal of Sociology, 96*(3), pp. 626-654.
[44] A. Rapoport (1997) "Order of Play in Strategically Equivalent Games in Extensive Form," *International Journal of Game Theory, 26*(1), pp.113-36.
[45] P. Resnick , K. Kuwabara, R. Zeckhauser , E. Friedman (2000a) “Reputation Systems,” *Communications of the ACM, 43*(12), pp. 45-48.
[46] P. Resnick, R. Zeckhauser (2000b) “Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay’s Reputatoin System,” *Working Paper for the NBER Workshop on Empirical Studies of Electronic Commerce*.
[47] S. Ross (1995) *Stochastic Processes*. John Wiley & Sons.
[48] J. Rouchier, M. O’Connor, F. Bousquet (2001) “The Creation of a Reputation in an Artificial Society Organized by a Gift System.” *Journal of Artificial Societies and Social Simulations, 4*(2).
[49] J. Sabater, C. Sierra (2001) “REGRET: A reputation Model for Gregarious Societies,” *4th Workshop on Deception, Fraud and Trust in Agent Societies*, Montreal, Canada.
[50] H. Simon (1981) *The Sciences of the Artificial*. Cambridge, Massachusetts: MIT Press.
[51] G. Strang (1988) *Linear Algebra and its Applications.* San Diego: Harcourt Brace and Jovanovich Publishers.
[52] S. Tadelis (1999) “What’s in a Name? Reputation as a Tradeable Asset,” *American Economic Review, 89*(3), pp. 548-563.
[53] R. L. Trivers, (1971) "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology, 46*, pp. 35-57.
[54] M. Waldman, A. D. Rubin, L. F. Cranor (2000) “Publius: A Robust, Tamper-Evident, Censorship-Resistent Web Publishing System,” *Proc. 9th USENIX Security Symposium*.
[55] S. Wasserman, K. Faust (1994) *Social Network Analysis: Methods and Applications*. Cambridge University Press.
[56] B. Yu, M. P. Singh (2001) “Towards a Probabililstic Model of Distributed Reputation Management,” *4th Workshop on Deception, Fraud and Trust in Agent Societies*, Montreal, Canada.
[57] G. Zacharia, P. Maes (1999) “Collaborative Reputation Mechanisms in Electronic Marketplaces.” *Proc. 32nd Hawaii International Conf on System Sciences*.
[58] P. R. Zimmerman (1995) *The Official PGP User's Guide*, Cambridge, Massachusetts: MIT Press.