Macau: A Basis for Evaluating Reputation Systems

Christopher J. Hazard
Hazardous Software
8929 Taymouth Ct.
Raleigh, NC 27613, USA
cjhazard@hazardoussoftware.com

Munindar P. Singh
Department of Computer Science
North Carolina State University
Raleigh, NC 27695-8206, USA
singh@ncsu.edu

Abstract

Reputation is a crucial concept in dynamic multiagent environments. Despite the large body of work on reputation systems, no metrics exist to directly and quantitatively evaluate and compare them. We present a common conceptual interface for reputation systems and a set of four *measurable* desiderata that are broadly applicable across multiple domains. These desiderata employ concepts from dynamical systems theory to measure how a reputation system reacts to a strategic agent attempting to maximize its own utility. We study a diverse set of well-known reputation models from the literature in a moral hazard setting and identify a rich variety of characteristics that they support.

1 Introduction

An agent's reputation summarizes information about it. Reputation systems are popular primarily because there are strong intuitive connections between an agent's reputation and both the utility it obtains and the utility another agent obtains when interacting with it.

Currently, reputation systems are evaluated in an ad hoc manner. Usually researchers compare a performance measure, often utility, of agents under a specific set of defined attacks for each reputation system, e.g., [Kerr and Cohen, 2009]. Alternatively, researchers use the ART testbed [Fullam *et al.*, 2005] to determine which agents (using different reputation systems prevail). However, ART does not compare reputation systems purely because outcomes depend on how well agents model their interactions and the environment. Further, it does not align incentives between obtaining a good reputation and increased utility [Krupa *et al.*, 2009].

In contrast, we propose a systematic approach for evaluation. Reputation is effective in enabling trust if agents' reputations change over time to increase predictive accuracy and to incentivize agents to cooperate. A rational agent only builds and maintains a positive reputation if doing so maximizes utility. Therefore, we approach reputation from a dynamical systems perspective, considering how an agent's behavior affects its reputation and vice versa. In this manner, we can evaluate a reputation system not in terms of some ad hoc attacks but based on how well it serves the purpose of inducing cooperative behavior over time.

We make three contributions. One, a set of quantifiable desiderata (monotonicity, accuracy, convergence, and unambiguity—hence the name *Macau*) by which to compare reputation systems in terms of their dynamical properties for a given environment, which together characterize their robustness to a wide range of attacks. Two, a uniform conceptual interface for reputation systems. Three, an approach for comparing reputation systems by examining their dynamics via how they influence and measure a rational agent's behavior. We demonstrate Macau on well-known reputation systems.

2 Reputation Model and Dynamics

We distinguish two roles: a *rater* evaluates a *target*. An agent may take on both roles in different interactions. We demonstrate our approach via the exchange of *favors*, e.g., the delivery of an item by a seller followed by payment by the buyer. The target decides whether to provide a favor and the rater rates the target. We present a common conceptual interface for reputation systems that treats them as black boxes. This interface consists of two functions reflecting the two fundamental features of a reputation system:

GetNext, a function that returns a target's new reputation after the target performs a specified action; and

GetPayoff, a function that returns the reward that a target can expect for a specified action given its reputation.

We represent the attributes of an agent (such as its utility functions, valuations, and discount factors) as its *type* $\theta \in \Theta$. We make no specific assumptions about the set of all possible agent types Θ and use θ as a label on appropriate functions. We treat reputation as an individual rater's belief of a target's type (measured in terms of the rater's utility) as $r \in R$, where R is defined by the given reputation system.

The target chooses how to behave, taking its current reputation into account. This behavior causes the rater to assess the target a certain way and (based on the reputation system) adjust the reputation of the target. Hence, the target's reputation would be mapped from its pre-action value, r, to its post-action value, r', based on the target's type, θ, the parameters of the interaction, $g \in G$, the environment, $\psi \in \Psi$, and the reputation system, $\xi \in \Xi$. The function $\Omega : \Theta \times G \times \Psi \times \Xi \times R \mapsto R$ captures the above intuitions.

Figure 1 shows a "cobweb" diagram from dynamical systems theory [Devaney, 1992]. We focus on real scalar reputations, ranging from $\underline{R}$ (worst) to $\overline{R}$ (best). $\underline{R}$ and $\overline{R}$ depend on the system and need not be finite. The diagonal represents unchanging reputation: a fixpoint exists wherever an Ω_θ function intersects the diagonal.

Figure 1 illustrates how reputation changes over time. Suppose a target has a bad reputation r_1. That is, a rater believes the target will likely behave in an undesirable fashion. The target's subsequent reputation (after performing its next action based on its strategy) is r_2, the value on the dashed line above r_1. (If the best strategy for the rater is to not interact with the target, then $r_2 = r_1$; hence, r_1 would be a fixpoint, as defined below.) The next time the target begins from r_2 and obtains an updated reputation of r_3. Each successive reputation is found by moving horizontally to the diagonal and then moving vertically to the new location on the dashed line. Here, the reputation converges to the (only one here) fixpoint marked by χ on each axis. If the target's reputation becomes higher (e.g., $\overline{R}$) than the fixpoint, it would "expend" some of its reputation, e.g., by providing poor service. As a result, its reputation would be lowered below the fixpoint. However, once the reputation is below the fixpoint, the target would behave nicely and rebuild its reputation back up to the fixpoint. Then it would expend it again, and so on.

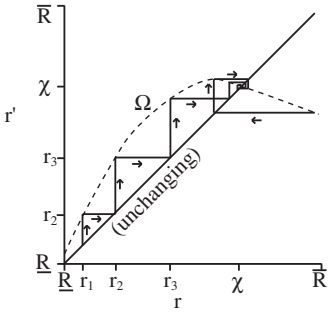

Figure 1: Reputation dynamics. X: current; Y: next.

An agent's behavior depends on whom it faces. Agent a might not provide a favor to agent b if a believes that b is not patient enough to return a favor to sustain a mutually beneficial long-term relationship [Hazard, 2008]. Here, a is not cooperative—but only because b is impatient. Patience coheres with reputation, because reputation helps select agents based on their expected future behavior. The rater's patience is the reason why reputation makes sense—ratings predict future payoffs to the rater and higher ratings promote future payoffs to the target. An impatient rater won't value a target's future good behavior and so won't return a favor; a target anticipating a defection won't cooperate.

Therefore, to enable appropriate comparisons, we consider a target facing an *ideally patient strategic* (*IPS*) agent, one that is indifferent to the time of when a specific utility change will occur. Thus, we remove the impact of the rater's impatience on the target's behavior. An IPS agent's discount factor γ approaches 1 indicating its time horizon is infinity.

Definition 1 *We define an* IPS *agent, b, as having an infinite time horizon such that b maximizes its average expected total utility, $E(\overline{U}_b(\theta_a))$, as a function of any agent a's type, θ_a, as the time horizon, represented as*

$$E(\overline{U}_b(\theta_a)) = \max_{\sigma_b} \lim_{\tau\to\infty} \frac{1}{\tau} \sum_{t=0}^{\tau} u_{\theta_b}(\theta_a, \sigma_{b,t}). \qquad (1)$$

We can now summarize our conception as follows. The favor exchange of a target with an IPS forms the environment in which the target exists. The Ω function depends on the target's type plus the given (implicit) reputation system, and maps a reputation to a reputation. It incorporates the IPS as the "standard" rater. Note that the IPS is an idealization to help us compare different reputation systems (and the targets within each reputation system) with respect to a baseline. Section 6 discusses some alternatives.

A fixpoint of a function is where its output equals its input. The properties of fixpoints, such as whether and how they attract or repel, govern the dynamics of feedback systems [Devaney, 1992]. In a desirable system, an agent's reputation should follow its type, i.e., its reputation should arrive at a fixpoint corresponding to its type.

The set of fixpoints of Ω_θ is $\{r \in R : r = \Omega_\theta(r)\}$. We define the function χ, which yields the stable fixpoint, if one exists, of a reputation system for a target of type θ, as $\chi(\theta) = \lim_{n\to\infty} \Omega_\theta^n(r_{\text{initial}})$, where Ω_θ^n is Ω_θ iterated n times. $\chi(\theta)$ depends on r_{initial}, the a priori belief that a rater has of a target, which is explicitly defined in some systems, or may be assumed to be the expected value over the distribution of possible reputations. However, the raters may have differing a priori beliefs or have misinformation about the targets, leading to differing initial reputations.

3 The Macau Reputation System Desiderata

3.1 Monotonicity

Informally, better agents should earn higher reputations. Here, "better" is judged from the perspective of the IPS rater. If, to an IPS rater c, target a's type is preferable to target b's type, then a's asymptotic reputation should be greater than b's reputation. In a monotonic reputation system, the expected utility an IPS agent would receive from a target can be predicted by comparing reputation values. That is, the reputation system successfully predicts the utility gained from a target.

We write a rater b's utility of entering an interaction with a as $u(\chi(\theta_a))$. Here, u is the payoff function that yields the value of a single transaction for a given reputation, which is a property of the reputation system.

Desideratum 1 MONOTONICITY*: A reputation system is* ***monotonic*** *if* $\forall\theta_a, \theta_b \in \Theta : E(\overline{U}_c(\theta_a)) \geq E(\overline{U}_c(\theta_b)) \Rightarrow u(\chi(\theta_a)) \geq u(\chi(\theta_b))$.

However, if c is indifferent across all agent types, that is, $\forall\theta_a, \theta_b \in \Theta : E(\overline{U}_c(\theta_a)) = E(\overline{U}_c(\theta_b))$, then the reputation system is **nondiscriminatory**, a generally undesirable subset of the otherwise desirable monotonic property.

3.2 Unambiguity

Reputation should be asymptotically unambiguous, meaning a target's eventual reputation should be independent of the rater's initial beliefs about it. Consider the fixpoints for a target's type in a reputation system. The ideal number is one.

When targets' reputations are unbounded and the mechanism has no fixpoints, all targets could end up with an unboundedly growing reputation, as in *saturating* in Figure 2. If, for all target types, Ω is strictly below the diagonal (except

at $\underline{R}$), as shown by *dissipating* in Figure 2, all targets would eventually end up with the worst possible reputation, $\underline{R}$.

If multiple fixpoints exist, the reputation is ambiguous: the fixpoint achieved depends on the rater's initial beliefs and possibly on rating errors. Consider *separating* in Figure 2. If the target's reputation is above the middle of the reputation domain then the target's reputation converges to χ. If the target's reputation starts below the middle, then it continually receives a lower reputation until it reaches the lowest possible value. The outcome depends solely on the initial reputation, which may be arbitrary, and yields a self-fulfilling prophecy.

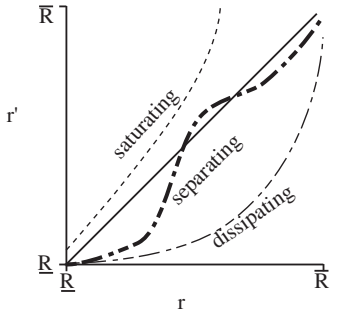

Figure 2: Meaningless reputation values.

Desideratum 2 UNAMBIGUITY*: A target's reputation should be unique, i.e.,* $\forall\theta \in \Theta : |\{r \in R : r = \Omega_\theta(r)\}| = 1$.

In some cases, e.g., when there are multiple contiguous fixpoints, ambiguity could be removed by abstracting such contiguous blocks of reputation values to a single value.

3.3 Accuracy

Accuracy represents a system's resilience to misinformation. If an agent's reputation is (significantly) incorrect, a system with good accuracy will quickly reduce the error. Figure 3 shows two systems, with the same fixpoint and the same derivative there. In *fast gain, slow expend* (FGSE), targets with low reputations quickly improve their reputation—meaning the line is much above the diagonal. However, reputations in FGSE can overshoot χ and oscillate near χ. In *slow gain, fast expend* (SGFE), targets gain reputation slowly, and targets that exceed the fixpoint would quickly expend a significant amount of reputation.

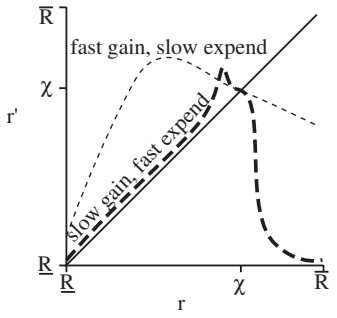

Figure 3: Reputation systems with different error amounts.

Thus, (this specific) FGSE is closer than (this specific) SGFE to the horizontal line at χ. Qualitatively, FGSE is preferable to SGFE because it is more stable and accurate. The ARME provides a quantitative comparison, yielding a lower error for FGSE. In general, the closer Ω is to a horizontal line ($r' = \chi$) for one-dimensional reputation measures, the lower the error is between the target's current reputation and its fixpoint:

Definition 2 *The* reputation measurement error (RME; average ARME)*,* $\epsilon \in [0, 1]$*, at reputation* r *for a target of type* θ *is the distance between a new reputation* $\Omega_\theta(r)$ *and the asymptotic reputation* χ*, normalized to* $|\overline{R} - \underline{R}|$*, the maximum distance between any two reputations.*

Desideratum 3 ACCURACY*: A system should yield low ARME,* $E(\epsilon)$*, with respect to the believed distribution of target types, represented by the probability density function* $f(\theta)$*, where* $E(\epsilon) = \int_\Theta f(\theta) \cdot E(\epsilon_\theta) d\theta$.

3.4 Convergence

Reputations should converge quickly, and be stable. Instead of the speed of computation which depends on the agents, we consider the dynamical properties of a reputation system. Convergence supports speed via robustness against errors, e.g., approximations in determining the correct reputation, because of limited abilities to observe and compute.

A fixpoint is *attracting* if the system asymptotically converges to the fixpoint when starting near it. A fixpoint is *repelling*, as in *self-affirming* in Figure 4, provided the dynamical system diverges from it unless exactly at it.

Accuracy increases near an attracting and decreases near a repelling fixpoint. Below, $d(r)$, the derivative magnitude at r, is $||\nabla\Omega(r)||_\infty$.

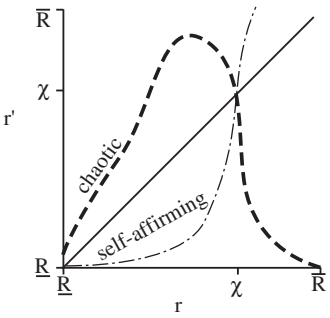

Figure 4: Large derivative magnitudes at fixpoint.

Desideratum 4 CONVERGENCE*: At the fixpoint,* $\chi(\theta)$*, the sequence of utility maximizing reputation values must be attracting and should converge quickly, that is,* $d(\chi(\theta)) < 1$*. All else equal, one reputation system is preferred to another if its* $d(\chi(\theta))$ *is smaller.*

When a target's reputation oscillates around a fixpoint, the reputation system is said to be *Lyapunov stable*, and we can treat it as being approximately convergent.

4 The Macau Evaluation Methodology

We now apply our desiderata on important reputation systems selected based on how clearly they measure reputation explicitly and provide implementations we can reconstruct.

4.1 Environment and Setup

In each round of our interaction model, a target begins with a specified reputation. It decides whether to offer a favor to a rater. If the target offers the favor, it incurs a cost of c to itself and the rater receives a benefit of b. The roles are then reversed, the other agent chooses whether to provide a favor with the same payoffs. To ensure gains from trade are possible, we examine these variables in the (partially overlapping) ranges of $c \in [1, 12]$ and $b \in [10, 30]$.

A discount factor captures how quickly an agent's utility falls for future events. Each agent multiplies the expected utility of a future event by γ^t, where $\gamma \in [0, 1]$ is the discount factor and t is the time of the event relative to the present. These future events are inputs to the agent's utility function to assess the expected utility of a given action or strategy.

In our simulations, we represent the target's strategies as a series of binary decisions, $\langle$favor, favor, nofavor, $\ldots\rangle$. We

limit the length of the strategies via a parameter, DEPTH $= \lceil \log(1-0.95)/\log(\gamma) \rceil$, so 95% of the total utility over the infinite horizon is captured. The possible strategies form the regular expression $\{\text{favor}, \text{nofavor}\}^{\text{DEPTH}}$.

Algorithm 1 ComputeNextReputation(raterModel, target, targetReputation)

```
1: bestUtility ← −∞
2: nextReputation ← targetReputation
3: for all s ∈ {favor, nofavor}^DEPTH do
4:    ⟨util, r⟩ ← ComputeUtilityAndReputationFromStrategy(raterModel, target, s, targetReputation)
5:    if util > bestUtility then
6:       bestUtility ← util
7:       nextReputation ← r
8:    end if
9: end for
10: return nextReputation
```

To find the optimal strategy given a discount factor, Algorithm 1 computes the utility gained for each possible strategy of the entire tree of the extended form game. Algorithm 1 approximates $\Omega_\theta(r)$ up to DEPTH. The overall computation of this Markov decision process is exponential in the number of decisions followed. A target's future expected utility for a reputation r is expressed recursively as

$$U(r) = \max_\sigma \left(u(r,\sigma) + \gamma \cdot U(N(r,\sigma)) \right), \qquad (2)$$

where σ is the agent's action, $u(r,\sigma)$ is the utility it expects to obtain for a given time step, and $N(r,\sigma)$ is the agent's new reputation after it performs σ. The agent's action will be the one that maximizes utility for the current reputation, r, that is, the outermost σ. Line 5 in Algorithm 1 invokes the interface functions GetNext and GetPayoff of Section 2.

4.2 Methodology Example: Beta Model

We describe our methodology using the Beta model as an example. In the Beta model, raters quantize interactions into positive and negative experiences and use a beta distribution to indicate the probability distribution that a target will perform positively in the future. Given a number of positive interactions, α, and negative interactions, β, the expected probability that a future interaction will be positive is $\frac{\alpha}{\alpha+\beta}$, the mean value of the beta distribution. A target's reputation is its expected probability of yielding a positive interaction.

1. Determine the update function. A rating, r, consists of the pair of the total numbers of positive ($i_{P,r}$) and negative interactions ($i_{N,r}$). The update function, n, for the Beta model is $n(r,\sigma_t) = \langle i_{P,r} + \sigma_t, i_{N,r} + (1-\sigma_t) \rangle$, where σ_t yields 1 if it offers the favor, else 0.

2. Determine the payoff function. The belief of a positive outcome, $b_P(r) = \frac{i_{P,r}}{i_{P,r}+i_{N,r}+1}$. With the linear relationship between reputation and utility, a target with $b_P = 0.25$ would receive half the price for a good than would a target with a $b_P = 0.5$. The utility, u, of a target of type θ for a favor at time t, is given by $u(p_B, t, \theta) = \gamma_\theta^t \cdot b_P \cdot \text{BENEFIT}$.

3. Run Algorithm 1 over Domain of Reputations. We ran Algorithm 1 on each possible reputation with 10 observations, from 10 positive and 0 negative observations, through 0 positive and 10 negative observations (for models other than Beta, we divided the reputation space into 10–100 points), using a variety of cost, benefit, and discount parameters. Except when otherwise noted, we ran discount factors from 0.0 to 0.8 in 0.1 increments. Finally, the entire set of tests needs to be run with various values of BENEFIT and COST to determine how consistently the model behaves across the range of favor sizes. For these values, we chose several combinations across the domains of c and b as outlined above.

4. Evaluate MONOTONICITY. If the rater's expected utilities are nondecreasing with respect to discount factor, the reputation system is monotonic, as for the Beta model with a superlinear relationship between reputation and price. If the utilities are constant, as is the case with the Beta model with linear and sublinear relationships between reputation and price, the reputation system is nondiscriminatory. If the rater utilities ever decrease with respect to increasing discount factor, the system is nonmonotonic. Alternatively, if no meaningful asymptotic reputation exists, the reputation system cannot be evaluated with respect to monotonicity.

5. Evaluate UNAMBIGUITY. We find UNAMBIGUITY by first examining each pair of successive inputs, say r_i and r_{i+1}, to Algorithm 1 for a given agent type (discount factor) and environment (BENEFIT and COST). If the line defined by $r' = r$ is crossed by (or coincides with) any two successive reputation values r_j and r_{j+1} when plotted based on their inputs (r_i and r_{i+1}), the intersection is a fixpoint. If zero or multiple fixpoints exist (as discovered for different values of i), the system fails UNAMBIGUITY. Otherwise, we use this unique fixpoint in computing the next measures.

6. Evaluate ACCURACY. Knowing the fixpoint, it is straightforward to calculate ACCURACY by computing the ARME of each output of Algorithm 1 with respect to each agent type and environment.

7. Evaluate CONVERGENCE. Knowing the fixpoint, the slope may be closely approximated by computing the slope of the line segment between the points immediately surrounding the fixpoint (or averaging the two nearby slopes if the fixpoint lies on the boundary between two line segments).

5 Applying Macau: Empirical Results

Here we discuss the results for each of the models we evaluate. We use our desiderata to compare reputation systems and find out how well they perform when faced with a strategic target agent. In doing so, we also validate that our desiderata are granular enough to distinguish differences between reputation systems, and that our results are intuitive. Table 1 summarizes our results discussed in the remainder of this section.

Results on the Beta Model

We refer to Jøsang's [1998] *Subjective* model, Teacy et al.'s [2006] *Travos* system, and Wang and Singh's [2007] *Certainty* model as the *Beta* model. These models differ in how they measure and aggregate uncertainty, but the underlying measurements are the same.

To evaluate the effect of uncertainty as measured by Travos and Certainty, we reduce the utility expected from agents of

Table 1: Summary of evaluation; the values listed are approximate averages across our experiments.

Reputation System	Unambiguity	Monotonicity	Convergence (defined so lower is better)	Accuracy (defined so lower is better)
Beta superlinear	yes	monotonic	0 and 0.9	0.4
Beta (sub)linear	yes	nondiscriminatory	0.9	0.45
Certainty	no	–	1	–
Discount Factor	yes	monotonic	< 0.1	0.02
Prob. Reciprocity	no	monotonic	no	0.2
Travos	yes	monotonic	0.8	0.2

uncertain trustworthiness. In the case of Travos, we multiply the expected utility by both the probability of a positive transaction and the certainty. For Certainty, we simply multiply the expected utility by the agent's belief value, as this accounts for both the probability of a positive transaction and the certainty. In both models, certainty is in the range of $[0, 1]$.

The results vary depending upon how we interpret an agent performing positively. Using a linear interpolation of the probability, a natural risk-neutral way of modeling utility, lead to results where no agents offered any favors and simply spent their reputations (thick line and circle line in Figure 5) and target reputations converged toward the minimum.

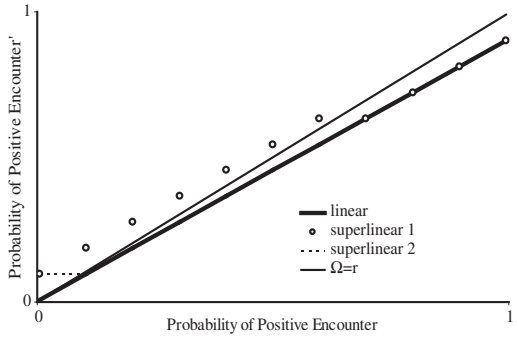

Figure 5: Beta model variants.

The sublinear results are the same as the linear. The Beta model fails MONOTONICITY, as all reputations end up the same—the reputation system cannot differentiate between them. In the superlinear case, that is, where a target is either risk-seeking or is not harmed as much by negative interactions, the Beta model fares quite well. The superlinear Beta model meets CONVERGENCE with positive slopes, either slowly with slopes of 0.9 or at the ideal of 0, and also meets MONOTONICITY by distinguishing higher values of discount factors. The Beta model's error in ACCURACY is mostly independent of the probability-utility relationship and ranges from 0.40 to 0.45. The optimal strategies against the Beta model are to repeatedly build up reputation over some period (depending on the parameters), and then exploit it.

Using Certainty's belief value instead of expected value yields results differing from the Beta model. Further, its characteristics become more pessimistic when evaluating against a group of three raters (who communicate to aggregate ratings) as opposed to an individual. In this network setting, one rational target interacts with three initially identical raters, all employing the same model. After each round of interactions, the raters exchange information about the target. The target's possible action space includes all combinations of actions and agents, so a target could conceivably act favorably to two agents and use its reputation to exploit a third.

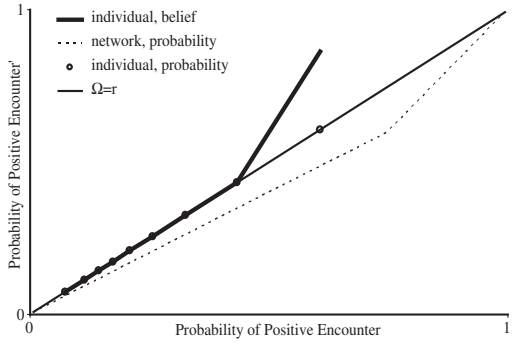

Figure 6: Certainty.

In Figure 6, *network, probability* is a typical case when a target is faced with a network of three raters. As shown by *individual, belief* and *individual, probability*, the targets are not incentivized to change their reputation until it exceeds a threshold, at which they always perform positively. Certainty meets neither UNAMBIGUITY nor MONOTONICITY, making it impossible to assess CONVERGENCE and ACCURACY.

Travos finds uncertainty by dividing the reputation space into five equal regions; finding the region containing the expected trustworthiness; and measuring certainty as the probability that the reputation is within it.

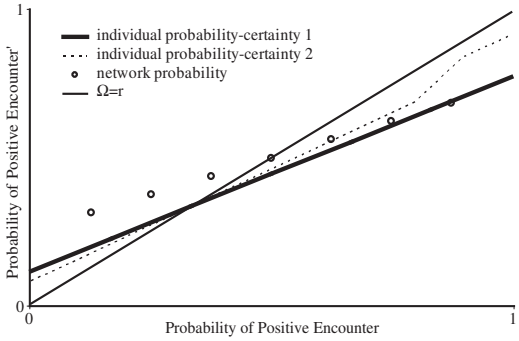

Figure 7: Travos.

Travos normalizes the magnitude of reputation information aggregated to a rater to prevent one rater from dominating another rater. However, doing so amplifies small numbers of observations. Using Travos's certainty as a multiplicative factor in utility, Travos meets MONOTONICITY, but is almost nondiscriminatory: most of the parameterizations yielded the same fixpoint, as Figure 7 shows.

Results on the Discount Factor Approaches

Here, agents strategically maximize utility while attempting to discover each others' discount factors [Smith and desJardins, 2009; Hazard and Singh, 2011]. An agent's discount factor measures its patience, weighting how it accounts for future utility by an exponentially decreasing function of time. The expected value of an agent's discount factor is its repu-

tation. An agent with a discount factor close to 0 would be myopic and greedy, whereas an agent with a discount factor close to 1 would offer favors if it expects the relationship (or reputation it earns from offering a favor) to be beneficial in the long run. As in Probabilistic Reciprocity [Sen, 2002], the reputation is explicitly connected with agents' utilities.

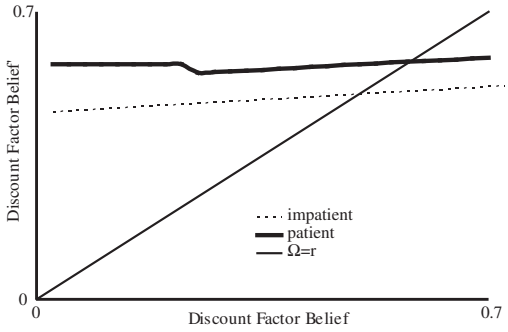

Figure 8: Discount Factor.

Across all parameterizations we examined, the results are similar to Figure 8, with all lines of the same shape, the main variation being the vertical location of the line. The agents strategically choose the optimal strategy that corresponds to their discount factors. Targets cannot credibly maintain an incorrect reputation, and their reputations converge quickly. MONOTONICITY is met: agents with a higher discount factor always offer better utility to a patient agent. UNAMBIGUITY is met: each agent type had exactly one fixpoint. The model fares well with CONVERGENCE, with the derivative at fixpoint being small and positive, usually less than 0.1. ACCURACY is high, with an error between 0.014 and 0.028.

6 Conclusions, Literature, Directions

Macau places a strategic target against an ideally patient rater and determines how a given reputation system induces the target to behave, leading to a particular trajectories of its reputation starting from different initial reputations. In this manner, Macau provides a principled, systematic basis for comparing reputation systems, measuring how they are designed and how well they would hold up against strategic attacks.

We have applied Macau on other systems, including Sen's [2002] *Probabilistic Reciprocity* and Zacharia and Maes' [2000] *Sporas* models. Kerr and Cohen's [2006] Trunits model requires a market where agents influence prices; it entails going beyond our simple favor scenario. Sierra and Debenham's [2005] information-theoretic model is geared toward richer interactions where agents have many possible actions.

Kerr and Cohen [2009] outline a number of possible ways that an agent could strategically improve its utility by being dishonest in a reputation system. Their "reputation lag" attack (a target alternates between honest and cheating periods) applies in a system that fails CONVERGENCE, where a target can exploit oscillations of its reputation. Similarly, their "value imbalance" (a target is honest with low-cost goods and dishonest with high-cost goods) and "reentry" (an agent continually opens new accounts to dishonestly use a new untainted reputation) attacks indicate a reputation system with poor ACCURACY, which does not recognize dishonest targets quickly. Our results are consistent with those of Kerr and Cohen; they find that the Travos, Beta, or Certainty models all can be effectively exploited by various strategies.

Collusion, side-payments, and Sybil attacks (using many pseudonyms to boost or reset reputation) are situations when agents may appear to not be individually rational. However, our desiderata can be adapted to reputation dynamics given a certain number of colluding raters attempting to boost the reputation of one scamming agent. To extend these desiderata, the colluding raters should be treated as one agent in terms of utility. The reputation of the scamming rater, that is, the colluding agent with the reputation inflated by the other colluding raters, can then be used directly in the desiderata.

Resnick and Sami [2007; 2008] use an information-theoretic approach to derive bounds on the damage an agent can wreak. Their method, as of Salehi-Abari and White [2009], limits an agent's influence. Although resistance to manipulation is not an explicit desideratum in our approach, it can be captured in dynamical terms. In order to manipulate its reputation, a target must perform actions so that a rater rates it highly, based on which the target would exploit it. This would be possible in a system that supported poor convergence or at least poor accuracy. The model proposed by Resnick and Sami does not account for rational agents that incorporate future rewards in their strategy, but instead focuses on Sybil attacks using randomized actions. Conversely, our desiderata focus on temporally strategic agents.

A variation on our measures would be to use a strategically malicious agent, whose utility is a function of other agents' loss. Such threats correspond to actors such as terrorists and angry customers, who willfully suffer loss to harm others.

The biggest weakness of the Macau approach is the computational complexity required to model reputation aggregation across a large number of agents and against strategic agents with high discount factors (hence long horizons). A future direction is developing approximate formulations that might produce useful results with lower complexity.

An important aspect, both strength and limitation, of Macau is that it considers ideally patient raters as the standard against which targets play, and which then leads to an assessment of reputation systems. Doing so yields idealized results but which may not apply when the rater is a human. Because human raters are often far from rational in the economic sense, it would be important to consider more realistic models of human behavior, such as prospect theory [Kahneman, 2011] or quantal response [McKelvey and Palfrey, 1995], as a way to judge the effectiveness of reputation systems where humans are directly the raters and consumers of reputation.

Interestingly, given enough data, Macau can potentially be used to mine the discount factors of raters. This could be used as a way of judging how far users deviate from the ideal.

Acknowledgments

This work was partially supported by the Army Research Laboratory in its Network Sciences Collaborative Technology Alliance (NS-CTA) under Cooperative Agreement Number W911NF-09-2-0053. We thank the anonymous reviewers for helpful comments.

References

[Devaney, 1992] Robert L. Devaney. *A First Course in Chaotic Dynamical Systems: Theory and Experiment*. Westview Press, Boulder, Colorado, October 1992.

[Fullam *et al.*, 2005] Karen Fullam, Tomas B. Klos, Guillaume Muller, Jordi Sabater, Andreas Schlosser, Zvi Topol, K. Suzanne Barber, Jeffrey S. Rosenschein, Laurent Vercouter, and Marco Voss. A specification of the agent reputation and trust (ART) testbed: Experimentation and competition for trust in agent societies. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 512–518, Utrecht, The Netherlands, July 2005. ACM Press.

[Hazard and Singh, 2011] Christopher J. Hazard and Munindar P. Singh. Intertemporal discount factors as a measure of trustworthiness in electronic commerce. *IEEE Transactions on Knowledge and Data Engineering*, 23(5):699–712, May 2011.

[Hazard, 2008] Christopher J. Hazard. ¿Por favor? Favor reciprocation when agents have private discounting. In *AAAI Workshop on Coordination, Organizations, Institutions and Norms (COIN)*, pages 9–16, Chicago, Illinois, July 2008.

[Jøsang, 1998] Audun Jøsang. A subjective metric of authentication. In *Proceedings of the 5th European Symposium on Research in Computer Security (ESORICS)*, volume 1485 of *LNCS*, pages 329–344, Louvain-la-Neuve, Belgium, 1998. Springer.

[Kahneman, 2011] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

[Kerr and Cohen, 2006] Reid Kerr and Robin Cohen. Modeling trust using transactional, numerical units. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust*, pages 1–11, New York, NY, USA, 2006. ACM.

[Kerr and Cohen, 2009] Reid Kerr and Robin Cohen. Smart cheaters do prosper: defeating trust and reputation systems. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 993–1000, Budapest, 2009. IFAAMAS.

[Krupa *et al.*, 2009] Yann Krupa, Jomi F. Hübner, and Laurent Vercouter. Extending the comparison efficiency of the ART testbed. In *Proceedings of the First International Conference on Reputation - Theory and Technology*, pages 186–199, Gargonza, Italy, 2009.

[McKelvey and Palfrey, 1995] Richard McKelvey and Thomas Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, July 1995.

[Resnick and Sami, 2007] Paul Resnick and Rahul Sami. The influence limiter: Provably manipulation resistant recommender systems. In *Proceedings of the ACM Conference on Recommender Systems*, pages 25–32, Minneapolis, MN, 2007.

[Resnick and Sami, 2008] Paul Resnick and Rahul Sami. The information cost of manipulationresistance in recommender systems. In *Proceedings of the ACM Conference on Recommender Systems*, pages 147–154, Lausanne, Switzerland, 2008.

[Salehi-Abari and White, 2009] Amirali Salehi-Abari and Tony White. Towards con-resistant trust models for distributed agent systems. In *International Joint Conference on Artificial Intelligence*, pages 272–277, 2009.

[Sen, 2002] Sandip Sen. Believing others: Pros and cons. *Artificial Intelligence*, 142(2):179–203, December 2002.

[Sierra and Debenham, 2005] Carles Sierra and John Debenham. An information-based model for trust. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 497–504, New York, NY, USA, 2005. ACM.

[Smith and desJardins, 2009] Michael J. Smith and Marie desJardins. Learning to trust in the competence and commitment of agents. *Autonomous Agents and Multi-Agent Systems*, 18(1):36–82, February 2009.

[Teacy *et al.*, 2006] W. T. Teacy, Jigar Patel, Nicholas R. Jennings, and Michael Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.

[Wang and Singh, 2007] Yonghong Wang and Munindar P. Singh. Formal trust model for multiagent systems. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1551–1556, Hyderabad, India, 2007.

[Zacharia and Maes, 2000] Giorgos Zacharia and Pattie Maes. Trust management through reputation mechanisms. *Applied Artificial Intelligence*, 14(9):881–907, October 2000.