A Framework for Decomposing Reputation in MAS into Competence and Integrity

Michael J. Smith
University of Maryland Baltimore County
Baltimore, MD 21250
msmith27@cs.umbc.edu

Marie desJardins
University of Maryland Baltimore County
Baltimore, MD 21250
mariedj@cs.umbc.edu

ABSTRACT

In multi-agent communities, trust is required when agents hold different beliefs or conflicting goals. We present a framework for decomposing agent reputation into *competence*—modeled as the probability of successfully carrying out an intended action—and *integrity*—modeled as a rational commitment to maintaining reputation. We demonstrate the usefulness of this approach in an iterated prisoner's dilemma (IPD) domain.

1. INTRODUCTION

This paper presents a framework for modeling components of trust and reputation: competence and integrity. The *competence* of an agent is its ability to correctly carry out its intended actions. *Integrity* is the commitment that an agent has to honor a stated commitment to take an action.

The decomposition of trust has been previously studied. Marsh treats competence similarly to our decomposition, but does not consider integrity [2]. McKnight and Chervany synthesized a high-level typology of trust, based on a broad survey of trust literature; competence and integrity are two of their primary categories of trust, along with benevolence and predictability [3]. In contrast, our research applies a formal framework founded on decision theory to explicitly model and separate competence and integrity

The ultimate goal of our research is to provide agents with a theoretical basis to learn about competence and integrity, and methods to make decisions based on this learned knowledge. In this paper, we show that agents with accurate estimates of the other agents' commitment to reputation (belief in the discount rate for the game) and competence can outperform strategies that do not model these factors.

2. APPROACH

In our framework, agents model competence as a simple probability of successfully completing a selected action. Integrity is modeled as a "commitment to reputation," which reflects that agent's belief about how long the game will last. This is currently modeled as a parameter, γ, which can be thought of as the discount rate for the game. A game's *true* γ is the probability that another turn will take place after the current iteration. The lower the perceived game γ, the more likely the temptation to cheat will overcome any advantage of long-term cooperation.

We tested our model using a variation of the iterated prisoner's dilemma (IPD) [1]. Each iteration (two-player game) has variable length, stochastically controlled by the true discount rate γ. The base payoffs are those of the "classic" IPD: the *reward* payoff R if both agents cooperate is 3; the *punishment* payoff P if both agents defect is 1; the *temptation* payoff T for an agent who defects when the other agent cooperates is 5; and the *sucker* payoff S for an agent who cooperates when the other agent defects is 0. In our variation, the base payoffs are multiplied by a *payoff multiplier* M. The payoff multiplier is generated randomly for each round of the game using an exponential distribution, which generates many relatively low-value transactions, but only rarely a high-value opportunity. The expected payoff multiplier, $\overline{M}$, is 1. This distribution creates the variation necessary for "confidence game" strategies, which can cooperate on low-value rounds and then cheat (defect) on high-value rounds to "cash in" on the high payoff.

Decision Strategies. We compared three baseline (classic) strategies and three trust-based (new) strategies. The baseline strategies are Always-Cooperate (ALL-C), Always-Defect (ALL-D), and Tit-for-Tat (TFT, which initially cooperates, then always matches an opponent's last move).

The three trust-based strategies have estimates of both player's competencies. The three strategies also incorporate, respectively, both agents' estimates of γ ("Both γ Both c" – BGBC), only their own γ ("Self γ Both c" – SGBC), and only the other agent's γ ("Other γ Both c" – OGBC). In these experiments, the estimates are always correct: the goal is to show that using accurate estimates of competence and integrity can improve performance.

Estimated Payoffs. The first step in determining payoffs in our framework is to compute the adjusted payoff matrix values. The estimated payoff for a given joint intent (i.e., intention to cooperate or defect for each agent) can be computed in a straightforward way by estimating the probability of each actual joint action from the competence and integrity estimates, then computing the expected payoff given the base payoffs and payoff multiplier.

Decision Making. The decision to cooperate or defect is based on a one-level recursive model of the other player's expected action. In particular, BGBC players using the trust-

AAMAS'05, July 25-29, 2005, Utrecht, Netherlands.

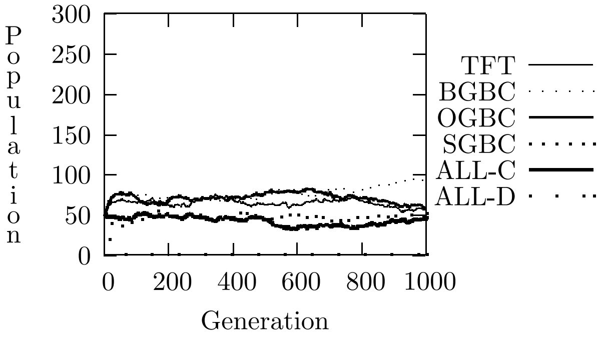

Figure 1: Competition with $\gamma = \hat{\gamma} = 0.9$, $c = \hat{c} = 1.0$.

based models will use the minimum of their γ and the other player's γ to estimate the game's true γ: either the game's probable length is short and one should defect, or the opponent thinks so, and one should likewise defect. Players who only model either their own γ (SGBC) or the other player's γ (OGBC) simply use that estimate as the true γ.

3. EXPERIMENTAL RESULTS

We performed a series of experiments in the IPD domain, using an evolutionary computing model to evaluate the alternate strategies. The fitness function is simply the average payoff for the games played on the most recent round. A selection bias ensures that individuals reproduce with a probability proportional to their fitness.

TFT and ALL-C were modeled by values of $\gamma = 1$; ALL-D was modeled by $\gamma = 0$. It can be argued that TFT is not perfectly modeled by $\gamma = 1$, since it does sometimes defect. However, given a cooperating opponent, TFT instantaneously reverts to $\gamma = 1$: it will never defect first, no matter how high the payoff. Note, however, that a TFT agent with competence lower than 1 *will* sometimes defect; modeling γ and c separately allows agents to differentiate unintended and intended defection.

In these experiments, the initial populations have 50 individuals per strategy; each game was run for 1000 generations; and agent performance was evaluated via a round-robin tournament. Twenty-five experiments were conducted, using a range of true discount factors and agent competence.

When the discount rate and competence are both high ($\gamma = 0.9, c = 1.0$, Figure 1), there is no advantage to any strategy, except that ALL-D consistently loses, becoming extinct within a few generations. By contrast, with high competence but a lower $\gamma = 0.8$ (Figure 2), the strategies that model at least the agent's own γ quickly eliminate all of the other strategies. Interestingly, OGBC, which models the other agent's γ but not its own, performs poorly and eventually dies out, along with TFT, ALL-C and ALL-D.

The trust-based strategies perform quite well in the high-competence (noise-free) environments. As γ decreases, the trust-based strategies outperform the standard strategies, including TFT. However, when we decrease competence, the trust-based strategies perform less well. Figure 3 shows that keeping $\gamma = 0.8$ but lowering competence to 0.8 results in a resounding success for TFT: only BGBC and SGBC manage to survive, and then only in low numbers.

The source of TFT's success when c is lower than γ is a weakness in the decision framework, namely, the assumption that a player who is defected upon will become "grim" and defect for the remainder of play. In fact, TFT and many other strategies have components of forgiveness (or forgetfulness), an aspect of trust that has been noted by other researchers [1] [4].

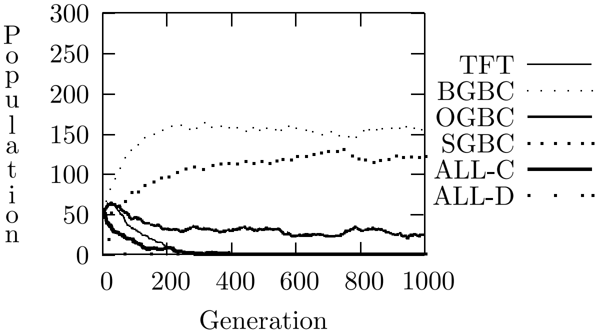

Figure 2: Competition with $\gamma = \hat{\gamma} = 0.8$, $c = \hat{c} = 1.0$.

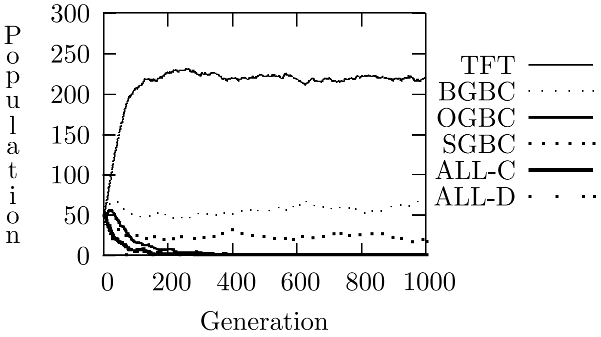

Figure 3: Competition with $\gamma = \hat{\gamma} = 0.8$, $c = \hat{c} = 0.8$.

4. CONCLUSIONS

The motivation for well grounded models of trust and reputation is increasing as multi-agent environments become more common and larger-scale. Trust and reputation problems have real-world, commercial analogs, such as e-commerce, contracting, and supply-chain management.

The experiments presented in this paper demonstrated that an explicit framework based on decision theory, combined with a framework for separating competence from integrity, can be effective in many situations. Future work includes explicitly modeling agents' attitudes towards forgiveness and forgetfulness, and exploring the theoretical implications of the recursive modeling of agents' beliefs. Our ultimate goal is to develop an effective means to learn integrity and competence estimates, allowing agents to adapt to new environments and changing behavior of other agents.

5. REFERENCES

[1] R. Axelrod. *The Evolution of Cooperation.* Basic Books, New York, 1984.

[2] S. P. Marsh. *Formalising Trust as a Computational Concept.* PhD thesis, University of Stirling, Apr. 1994.

[3] D. H. McKnight and N. L. Chervany. Trust and distrust definitions: One bite at a time. In *Proceedings of Trust in Cyber-societies*, pages 27–54, London, UK, 2000. Springer-Verlag.

[4] J. Sabater and C. Sierra. Regret: Reputation in gregarious societies. In *Proc. of the 5th Intl. Conf. on Autonomous Agents*, pages 194–195, Montreal, 2001.