Towards Incentive-Compatible Reputation Management

Radu Jurca and Boi Faltings

Artificial Intelligence Laboratory (LIA),
Computer Science Department, Swiss Federal Institute of Technology (EPFL)
CH-1015 Ecublens, Switzerland
{radu.jurca, boi.faltings}@epfl.ch
http://liawww.epfl.ch/

Abstract. Traditional centralized approaches to security are difficult to apply to large, distributed, multi-agent systems. Developing a notion of *trust* that is based on the *reputation* of agents can provide a softer notion of security that is sufficient for many MAS applications. However, designing a reliable and "trustworthy" reputation mechanism is not a trivial problem. In this paper, we address the issue of incentive-compatibility, i.e. why should agents report reputation information and why should they report it truthfully. By introducing a side-payment scheme organized through a set of broker agents we make it rational for software agents to truthfully share the reputation information they have acquired in their past experience. The theoretical results obtained were verified by a simple simulation. We conclude by making an analysis of the robustness of the system in the presence of an increasing percentage of lying agents.

Keywords: trust, reputation mechanism, incentive-compatibility.

1 Introduction

Software agents are a new and promising paradigm for open, distributed information systems. However, besides the many practical solutions this new paradigm provides, it also brings along a whole new set of unsolved questions. One of the issues that has attracted a lot of attention lately is security. Traditional, centralized approaches of security do no longer cope with the challenges arising from an open environment with distributed ownership in which agents inter-operate. [7], [4], [5]

We focus in particular on the problem of trust, i.e. deciding whether another agent encountered in the network can be trusted, for example in a business transaction. In closed environments, trust is usually managed by authentication schemes that define what agents are to be trusted for a particular transaction. In an open environment, fixed classifications must be replaced by dynamic decisions. One important factor in such decisions is an agent's *reputation*, defined as information about its past behavior.

The most reliable reputation information can be derived from an agent's own experience. However, much more data becomes available when reputation information is shared among an agent community. Such mechanisms have been proposed and also practically implemented. The various rating services on the internet are examples of such mechanisms.

It is however not at all clear that it is in the best interest of an agent to truthfully report reputation information:

- by reporting any reputation information, it provides a competitive advantage to others, so it is not in its interest to report anything at all.
- by reporting positive ratings, the agent slightly decreases its own reputation with respect to the average of other agents, so it is a disadvantage to report them truthfully.
- by reporting fake negative ratings, the agent can increase its own reputation with respect to others, so it is an advantage to report them falsely.

Thus, it is interesting to consider how to make a reputation mechanism *incentive-compatible*, i.e. how to ensure that it is in the best interest of a rational agent to actually report reputation information truthfully. This is the problem we address in this research.

2 An Example of an Incentive-Compatible Mechanism

As the first step in our research, we have constructed an example of a reputation sharing mechanism that is indeed incentive-compatible, thus showing that such a mechanism is possible. From the considerations given above, it is clear that an incentive-compatible mechanism should introduce side payments that make it rational for agents to truthfully share reputation information. In our mechanism, these side payments are organized through a set of broker agents, called R-agents, that buy and sell reputation information. We assume that no other side payments occur between any agents in the system.

As a first step, we show a mechanism which is incentive-compatible for a certain scenario under the condition that all other agents behave rationally, i.e. also report the truth. The problem of initialization is not studied for now, but rather, once that the system started, and assuming that in the system there is a majority of agents reporting the truth, we focus on a mechanism that also makes it in the best interest of agents to share information truthfully.

The scenario is the following. We assume we have *n* agents: a_i for $i = 1 \ldots N$, that interact pairwise in an iterated Prisoner's Dilemma environment. In each round, two agents together invest *I* units of money in an idealized business that pays of *f(I)* units with certainty, where *f* is some function. The agents can cooperate, and each invest an equal amount of money *(I/2)*, or can cheat and not invest anything. At the end of each round the benefits are split equally between the two partners, whether they have cheated or not. Each agent will cooperate with probability p_i, or defect with probability *1-* p_i.

Each agent can buy reputation information about another agent from an R-agent at a cost *F*, and later sell reputation information to any R-agent at a price *C*. Reputation is represented as a single real number in the interval $[0.0, 1.0]$. Agents report either 0 for a defection or 1 for cooperation, and the reputation r_i of an agent a_i is computed as the mean of all the reports about that agent:

$$r_i = \frac{\sum_{j=1}^{k} report_j}{N} \tag{1}$$

where $report_j$, $j = 1 \ldots k$ are the *k* reports that have been filed for agent a_i and can take the values 0 or 1.

In our scenario, agents systematically buy reputation information before engaging in business with another agent. Agents are only allowed to sell a report for an agent when they have previously bought reputation information for that agent.

To make the reputation mechanism incentive-compatible, we then have the following conditions:

1. Agents that behave as good citizens, i.e. report truthfully the result of every interaction with another agent, should not loose any money:
   E[F] $\leq$ *E[C|truthful report]*
2. Agents that report reputation incorrectly should gradually loose their money:
   E[F] $\geq$ *E[C|false report]*

To satisfy these conditions, we propose the following mechanism. The basic idea is that R-agents will only pay for reports if they match the next report filed by another agent. In order to prove the rationale behind this rule, suppose that we consider the reputation of agent a_i and let us compute the probabilities for the following events:

– agent a_i cooperates in two consecutive rounds: p_i^2
– agent a_i defects two consecutive rounds: $(1 - p_i)^2$
– agent a_i cooperates then defects: $p_i(1 - p_i)$
– agent a_i defects then cooperates: $p_i(1 - p_i)$

The probability that agent a_i behaves the same way in consecutive rounds is thus:

$(1 - p_i)^2 + p_i^2 = 1 - 2p_i + 2p_i^2$ which is bounded by $[0.5, 1]$.

On the other hand, the probability that agent a_i will change its behavior in two consecutive rounds is:

$2p_i(1 - p_i)$ which is bounded by $[0, 0.5]$.

Assuming that the other agents will report the truth, and that a_i will behave the same way on the next interaction, the optimal strategy for an agent is to report behavior truthfully, since this means it will be paid with probability of at least 0.5.

The remaining question is how much agents should be paid. For this, we need to consider that agents can only file a report if they actually did business with the agent, i.e. if they trusted the agent. Before each business begins, agents assess the trustworthiness of their partner. The business is done only if both partners agree.

The expected payoff an agent receives for a report on another agent a_i can be computed by analyzing the following situations:

a) the reputation of a_i is too low, which means that no business will be conducted and no report can be sold. In this case, the payoff is 0;
b) business is conducted, but the partner agent changes its behavior in the next round. Therefore, the agent's reporting will be considered as false. In this case, the payoff is also 0;
c) business is conducted and the partner agent behaves in the same way in the next round. The payoff is C in this case;

Therefore:

$$E[\textit{payoff}] = 0 \cdot Pr(\textit{case a}) + 0 \cdot Pr(\textit{case b}) + C \cdot Pr(\textit{case c})$$

We assume that an agent trusts agent a_i, and thus enters into business with a_i, whenever it expects it to yield a profit. Therefore q, the probability that an agent will trust another agent, is given as $q = Prob(Out > 0)$, where Out is the expected outcome of the business.

$$Out = \frac{1}{2}\left[(1 - p_i) \cdot f\left(\frac{I}{2}\right) + p_i \cdot f(I)\right] - \frac{I}{2}. \tag{2}$$

where $f(I)$ is the business payoff function for I units invested. Assuming a monotone increasing function f, the condition $Out > 0$ is equivalent to $p_i > \theta$, where θ is some constant that depends only on the business payoff function f. Therefore, $q = Pr(p_i > \theta)$.

The probability of conducting business is equal to the probability that both agents trust one another. Therefore:

Pr(case a) = $(1\text{-}q^2)$;
Pr(case b) = $2q^2 p_j(1 - p_j)$;
Pr(case c) = $q^2(1 - 2p_j + 2p_j^2)$;

for different p_j. Because agents are selected randomly with uniform probability to play the game, we can compute the mean value for the payoff:

$$E[\textit{payoff}] = C \cdot \frac{\sum_{j=1}^{N} q^2(1 - 2p_j + 2p_j^2)}{N} \tag{3}$$

There is a unique value for the price F of reputation information that would make the entire mechanism self-sustaining (i.e. R-agents neither loose nor win any money). The price of reputation information F:

$$F = E[\textit{payoff}] \tag{4}$$

However, in a practical implementation we can compute it simply as the moving average of the observed payoffs which must converge to an equilibrium value given by equation 4.

3 Testing Scenario

In this environment we propose the introducing of specialized "review agents" (R-agents) (the equivalent of professional survey companies) which are not allowed to play the game, but which have as a goal to obtain and sell information about the reputation of business agents. In the environment we will have several such agents, so that there is competition between them. One business agent will buy reputation information from one R-agent, but might get paid by all the R-agents. Therefore, we will divide the payoff C an agent receives for reporting correct reputation by the number of R-agents in the system, and agents will sell reputation information to all R-agents In our present work the fact that there are more R-agents makes no difference. Business agents randomly

select the R-agent from which they will buy reputation. However, in future work we will also implement a direct interaction derived reputation model of R-agents. Business agents will be able to develop preferences for R-agents that correctly provide reputation information. Another reason for the presence of more R-agents in the environment is system robustness.

We used a linear business payoff function in our experiments: $f(I) = x \cdot I$, where x is a coefficient greater than 1. By tuning x we modify the trading particularities of the environment: a small value for x corresponds to harsh trading environment where it is very important to trust your partner, while a big value for x would correspond to a friendly trading environment, where positive payoff is more probable, regardless of the partner's cooperation. An average value for x would correspond to a trading environment where agents make the decision of whether or not to conduct business with their partners by evaluating the inequality $pi > \theta = 0.5$. By replacing this in equation 2, to obtain the corresponding $\theta = 0.5$, we need to set $x = 1.33$.

The expected payoff for filing reputation reports would depend on the probabilities p_i of all the agents. The fact that these probabilities are unknown can be solved by using for the price F of buying reputation information the moving average of the payoffs obtained by the agents for selling reputation reports. The price F converges to the unique solution of the equation 4.

4 Experiments

The simulation of the above described environment shows encouraging results. We have used ten thousand business agents in our environment, and ten R-agents. The first test was to see whether the trust model implemented can help the trading between agents. Figure 1 shows the average wealth of cooperating and cheating agents. As it can be seen, the mechanisms implemented help cooperative agents to successfully detect and isolate cheating agents.

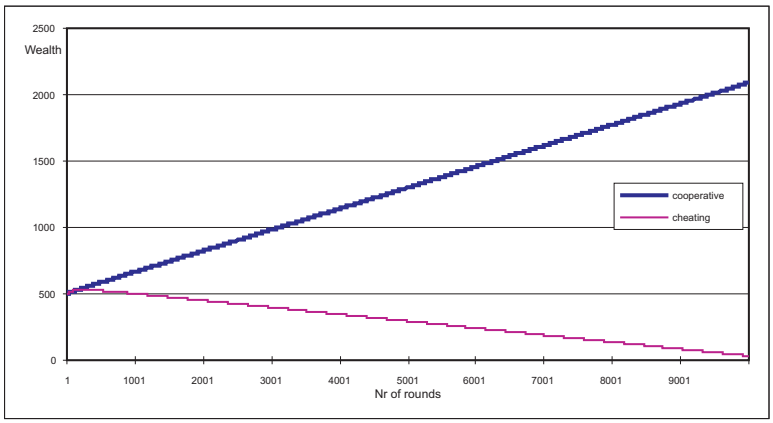


Fig. 1. Average wealth of cooperative and cheating agents.

For our next experiment, we tried to see if agents have an incentive to use the reputation information in their business. For that, we introduced in our society a percent of "lonely" agents that do not use the trust model. Figure 2 plots the average wealth of the "social" agents, who use the trust model, and the average wealth of the "lonely" agents against the number of rounds. Results show that social agents are better off than lonely agents.

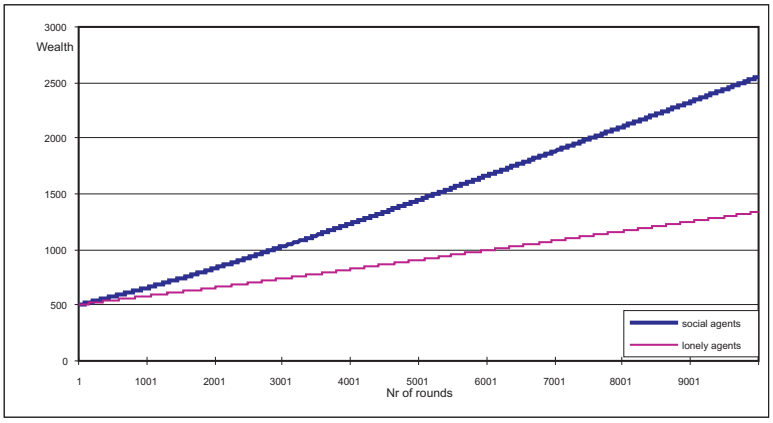


Fig. 2. Evolution of average wealth for lonely and social agents.

Finally, we were interested to see if agents have the incentive to report true reputation. For that, we introduced in our environment 1% of "lying" agents, i.e. agents that do not report the truth when they are asked. Figure 3 plots the evolution of the average wealth for truthful and lying agents. A more detailed analysis of the system's behavior in the presence of lying agents is presented in the following section.

These results allow us to believe that our model can be successfully used for providing the agents with the incentive to report true reputation information. In future work we will try to improve this model and find the combination of parameters that yields the best results.

5 Analysis of Mechanism Robustness in the Presence of Lying Agents

In Figure 3 we have seen satisfying results for the presence of 1% lying business agents in the system. In this section we will analyze the system's behavior as this percentage increases.

We will assume consistently lying agents (i.e. agents that lie all the time) adopting only one of the following three different strategies:

a) lying agents report the opposite of the observed behavior of the partner;

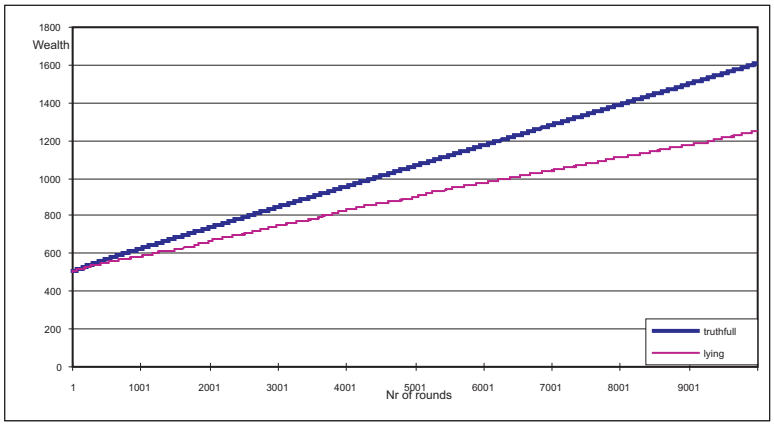


Fig. 3. Evolution of average wealth for truthful and lying agents (1%)

b) lying agents always report negative reputation for their partner;
c) lying agents give random reports for their partners. All lying agents lie according to the same strategy, and they do not change their strategy during the lifetime of the system.

All lying agents lie according to the same strategy, and they do not change their strategy during the lifetime of the system.

Let us denote by:

p_i the real reputation of agent a_i (i.e. the number of times the agent cooperated divided by the total number of interactions the agent had, aka. cooperation level of the agent);
p_i^t the perceived reputation of the agent a_i as known at time instance t by R-agents. t has the meaning of *number of proposed businesses*;
q the real percent of cooperative agents (i.e. the percent of agents whose cooperation level p_i is greater than a threshold θ);
q_t the percent of perceived cooperative agents (i.e. the percent of agents whose perceived reputation p_i^t is greater than a threshold θ);
α the percent of lying agents.

Assuming that $p_i^0 = 1$ (i.e. agents are initially considered trustworthy), the evolution of p_i^t in each of the three lying strategies enumerated above is given by the following equations:

$$p_i^{t+1} = \begin{cases} (1-\alpha)\frac{p_i^t t + p_i}{t+1} + \alpha\frac{p_i^t t + (1-p_i)}{t+1} = \frac{p_i^t t + (p_i - 2p_i\alpha + \alpha)}{t+1} & \text{for case (a)} \\ (1-\alpha)\frac{p_i^t t + p_i}{t+1} + \alpha\frac{p_i^t t + 0}{t+1} = \frac{p_i^t t + (p_i - p_i\alpha)}{t+1} & \text{for case (b)} \\ (1-\alpha)\frac{p_i^t t + p_i}{t+1} + \alpha\frac{p_i^t t + 0.5}{t+1} = \frac{p_i^t t + (p_i - p_i\alpha + 0.5\alpha)}{t+1} & \text{for case (c)} \end{cases} \quad (5)$$

The convergence value of p_i^t when t approaches ∞ depends only on the true cooperation level of that agent, p_i, and on the percent of lying agents, α. The equations

above also show the impact different lying behaviors have on the perceived reputation of the agents within the system.

In the first case, as the value of α increases from 0 to 0.5, the perceived reputation p_i^t is biased towards the value 0.5. Cooperative agents will have a slightly lower reputation, while defective agents will have a slightly better reputation. For $\alpha = 0.5$, the reputation information becomes completely useless because all agents will have a perceived reputation of 0.5. Moreover, as α grows bigger than 0.5, reputation information is misleading since defective agents are perceived as cooperative and cooperative agents are perceived as defective.

In the second case, as the value of α increases, the perceived reputation p_i^t of all agents converges to 0. The advantage over the previous case is that cooperative agents will always have higher perceived reputation than defective agents.

In the third case, as the value of α increases, the values for perceived reputation approach the value 0.5. However, cooperative agents will always have perceived reputation higher than 0.5, while defective agents will always have perceived reputation lower than 0.5. In this case the system will be able to build the most accurate reputation information since the error $|p_i^t - p_i|$ has the smallest increase with the increase of α.

The effect of the error introduced by lying agents in the perceived reputation of the business agents is reflected in the average increase of the wealth of the agents. Since on the average the reputation payments sum to zero (i.e. overall, the amount of money paid for retrieving reputation information is equal to the total amount of money received for filing reputation reports) we will consider only the wealth increase resulted from business between two agents.

Let us consider two agents a_i and a_j having the opportunity to do business. The probability that business is conducted, $Pr(business)$, is:

$$Pr(business) = Pr(p_i^t > \theta) \cdot Pr(p_j^t > \theta) \approx q_t^2$$

The expected payoff of this particular business opportunity is:

$$E[payoff] = q_t^2 \frac{(x-1)I}{2} (p_i + p_j) \tag{6}$$

where $f(I) = x \cdot I$ is the business payoff function, I is the proposed investment and p_i and p_j are the cooperation levels of the two agents. Because agents are chosen randomly, we can compute an average expected payoff as:

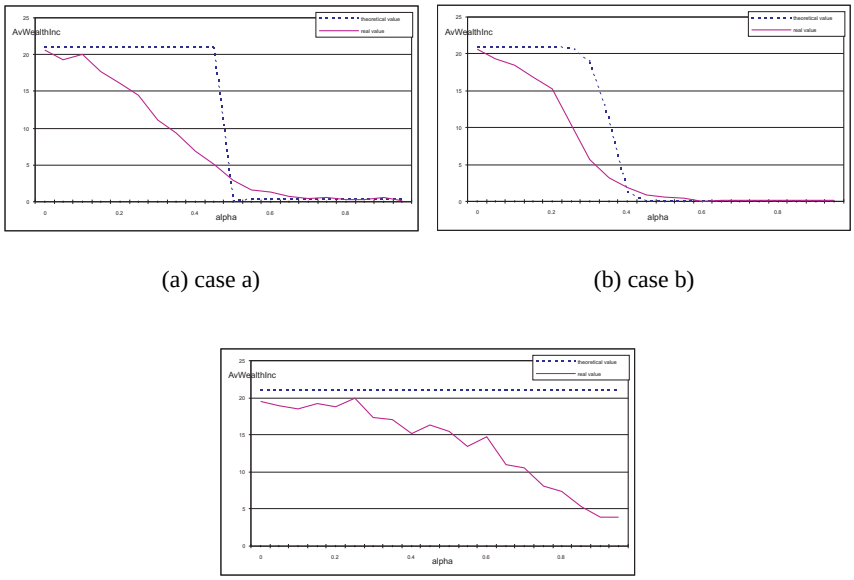
$$E[payoff] = q_\infty^2 \cdot (x-1) \cdot \overline{I} \cdot \overline{p} \tag{7}$$

where $\overline{p}$ is the average real cooperation level of the agents who are perceived as cooperative, $\overline{I}$ is the average investment, and $q_\infty = \lim_{t \to \infty} q_t$. Therefore, the average wealth increase for one business round will be:

$$AvWealthInc = \frac{E[payoff]}{N} \tag{8}$$

where N is the total number of agents in the system. The $AvWealthInc$ is always positive, and is affected by the presence of lying agents only through the values of q_∞ and $\overline{p}$. Knowing the distribution of the values of p_i, and considering the equations in 5,

6, 7 and 8, we can determine the theoretical dependence of $AvWealthInc$ on α. Figure 4 plots the theoretical dependence against the observed values in the simulation.



(a) case a)

(b) case b)

(c) case c)

Fig. 4. Average Wealth Increase depending on the percent of lying agents

6 Related Work

In [7] the authors present a definition of trust by identifying its constructs: *trust-related behavior, trusting intentions, trusting beliefs, institution-based trust* and *disposition to trust*. On the other hand, the Social Auditor Model, presented in [6], accounts for the process humans undergo when taking trusting decisions. Combining the two, a framework is obtained in which different trust and reputation models can be compared and classified. In the present paper, we present a simple trust model within this framework that uses only the *trusting beliefs* construct (the extent to which one believes that the other person has characteristics beneficial to one) from the definition in [7] under the name of reputation, and a simple decision process in which agents can take binary decisions (yes or no) about whether to interact or not with other agents. For simplicity, we also combined the four different aspects of reputation (competence, benevolence, integrity and predictability) into one number.

Mui et al. [8] present an extensive reputation typology classified by the means of collecting the reputation information. As stated before, we employ only two categories

from the typology in our trust model: the direct interaction-derived reputation and the propagated (from other agents) indirect reputation.

There are a number of systems that implement trust mechanisms based only on direct interaction-derived reputation: [1], [2], [6], [10], [3]. However, all these systems deal with an environment with a relatively small number of agents where direct reputation can be build. These models will not work in a very large environment because the time necessary for building direct reputation would be too large.

[9] proposes a solution that takes into consideration the reputation information reported by other agents. However, this solution we believe is not realistic because it does not provide any incentive for the agents to report the reputation information. Besides, each agent has to implement a rather complicated mechanism for judging the information it has received from its peers.

7 Conclusion

In our work, we built a successful trust model in an environment where a big number of trading agents conduct business. We have done so by using a reputation-based trust model in which both direct interaction-derived reputation and propagated indirect reputation is used. Special care was dedicated to the problem of incentive compatibility. By introducing a mechanism of payments, and a separation of goals through two kind of agents (business and review agents) we have shown that it is possible to make it in the best interest of the agents to share reputation information and to share it truthfully.

Acknowledgements

We thank Monique Calisti for her help and constructive remarks while writing this paper.

References

1. A. Birk. Boosting Cooperation by Evolving Trust. *Applied Artificial Intelligence*, 14:769–784, 2000.
2. A. Birk. Learning to Trust. In M. Singh R. Falcone and Y.-H. Tan, editors, *Trust in Cyber-societies*, volume LNAI 2246, pages 133–144. Springer-Verlag, Berlin Heidelberg, 2001.
3. A. Biswas, S. Sen, and S. Debnath. Limiting Deception in a Group of Social Agents. *Applied Artificial Intelligence*, 14:785–797, 2000.
4. L. Kagal, T. Finin, and J. Anupam. Moving from Security to Distributed Trust in Ubiquitos Computing Environment. *IEEE Computer*, December 2001.
5. L. Kagal, T. Finin, and J. Anupam. A Delegation-based Distributed Model for Multi Agent Systems System. http://www.csee.umbc.edu/˜finin/papers/aa02, 2002.
6. R. Kramer. Trust Rules for Trust Dilemmas: How Decision Makers Think and Act in the Shadow of Doubt. In M. Singh R. Falcone and Y.-H. Tan, editors, *Trust in Cyber-societies*, volume LNAI 2246, pages 9–26. Springer-Verlag, Berlin Heidelberg, 2001.
7. H. McKnight and N. Chervany. Trust and Distrust: One Bite at a Time. In M. Singh R. Falcone and Y.-H. Tan, editors, *Trust in Cyber-societies*, volume LNAI 2246, pages 27–54. Springer-Verlag, Berlin Heidelberg, 2001.

8. L. Mui, A. Halberstadt, and M. Mohtashemi. Notions of Reputation in Multi-Agents Systems:A Review. In *Proceedings of the AAMAS*, Bologna, Italy, 2002.
9. M. Schillo, P. Funk, and M. Rovatsos. Using Trust for Detecting Deceitful Agents in Artificial Societies. *Applied Artificial Intelligence*, 14:825–848, 2000.
10. M. Witkowski, A. Artikis, and J. Pitt. Experiments in building Experiential Trust in a Society of Objective-Trust Based Agents. In M. Singh R. Falcone and Y.-H. Tan, editors, *Trust in Cyber-societies*, volume LNAI 2246, pages 111–132. Springer-Verlag, Berlin Heidelberg, 2001.