An Adaptive Probabilistic Trust Model and its Evaluation (Short Paper)

Chung-Wei Hang
chang@ncsu.edu

Yonghong Wang
yhwang_y2k@yahoo.com

Munindar P. Singh
singh@ncsu.edu

Dept. of Computer Science
North Carolina State University
Raleigh, NC 27695-7535, USA

ABSTRACT

In open settings, the participants are autonomous and there is no central authority to ensure the felicity of their interactions. When agents interact in such settings, each relies upon being able to model the trustworthiness of the agents with whom it interacts. Fundamentally, such models must consider the past behavior of the other parties in order to predict their future behavior. Further, it is sensible for the agents to share information via referrals to trustworthy agents. Much progress has recently been made on probabilistic trust models including those that support the aggregation of information from multiple sources. However, current models do not support trust updates, leaving updates to be handled in an ad hoc manner.

This paper proposes a trust representation that combines probabilities and certainty (defined as a function of a probability-certainty density function). Further, it offers a trust update mechanism to estimate the trustworthiness of referrers. This paper describes a testbed that goes beyond existing testbeds to enable the evaluation of a composite probability-certainty model. It then evaluates the proposed trust model showing that the trust model can (a) estimate trustworthiness of damping and capricious agents correctly, (b) update trust values of referrers accurately, and (c) resolve the conflicts in referral networks by certainty discounting.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*multiagent systems*

General Terms

Algorithm, Experimentation

Keywords

Trust, Probabilistic Model, Adaptive Model

1. INTRODUCTION

Two key characteristics of open environments are that agents act independently, and may join and leave freely. Interactions among autonomous agents are based on a notion of trust. Here we take a narrow characterization of trust that considers it from a probabilistic standpoint. Trust in this view is a prediction of the quality of the future behavior of another party (whose trustworthiness is being evaluated).

Intuitively, agent A's trust of agent B can be estimated by a probability, namely, the probability of A getting a positive outcome from an interaction with B. However, a naive representation cannot distinguish getting 1 positive outcome out of 2 interactions, from getting 100 positive outcomes out of 200 interactions, because in both cases, the probability is equal to 0.5. Therefore, some trust models [4, 6] introduce certainty, interpreted as a confidence measure about the probability. Wang and Singh (W&S) motivate that certainty as applied in trust modeling should support following features [6]: (1) certainty increases as the number of interactions increases (if the ratio of positive and negative outcomes is fixed), and (2) certainty decreases as the conflict increases in the interaction experience (if the total number of interactions is fixed).

In open environments, an agent would often need to interact with a stranger, i.e., a party with whom the agent has had limited or no prior experience. The notion of *referrals* addresses this situation. An agent may ask its acquaintances about a stranger. An acquaintance may reply if it has prior experience with the stranger, or refer to others. This results in the so-called referral networks. To apply in referral networks, trust models should provide mechanisms for agents to aggregate trust values from multiple sources. The challenges of trust aggregation are (1) how to make use of the trust values provided by unreliable acquaintances, (2) how to combine the trust values from direct experience and referrals, and (3) how to update the trust placed in referrers based on the referrals they provide.

W&S propose a trust model based on a probability-certainty density function (PCDF) [6]. In their model, a trust value is composed of the probability of a positive outcome, and the certainty placed in the probability. They define two operators, aggregation and concatenation, for combining the trust values from multiple sources [5]. This paper extends W&S's model, providing an efficient algorithm for updating the trust values of agents based on the referrals they provide. We design an empirical evaluation testbed and perform experiments, which suggest that our model can identity good and bad agents efficiently, with limited prior experience. The experiments show that our model can adapt to agents' behavior changes over time, which is a common challenge in open systems.

2. RELATED WORK

In recent years, many trust models have been developed. However, a general trust evaluation testbed does not exist. The most critical reason is that the models instantiate different theories, taking in different inputs (e.g., number of experience, probability and confidence, and so on), generating different outputs (e.g., continuous versus discrete outputs), and exhibiting different standards (e.g.,

Cite as: An Adaptive Probabilistic Trust Model and its Evaluation (Short Paper), Chung-Wei Hang, Yonghong Wang and Munindar P. Singh, *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, Padgham, Parkes, Müller and Parsons (eds.), May, 12-16., 2008, Estoril, Portugal, pp. XXX-XXX.

different interpretations of good experience). This makes designing an ideal open trust evaluation testbed infeasible.

For instance, one of the famous trust evaluation testbeds, ART [1], aims to provide trust research community a unified platform for evaluation and competition. In ART, agents use trust strategies to exchange expertise with others to appraise paintings. Agents make money by providing accurate appraisals to receive more paintings to appraise in the next time step. However, agents' expertise is limited to appraising some of the paintings. They are required to exchange expertise with trustworthy others. They can also exchange trust values with others. The agent who has the most money at the end wins the game.

Gómez *et al.* [2] represent trust as the discrepancy between the information received and one's own experience. Trust values are collected from four kinds of information sources: direct experience, advertisement, recommendation, and global trust. Gómez *et al.* introduce two measures to handle certainty: intimacy (i.e., the number of experiences) and predictability (i.e., the dispersion of the data). They use ART to evaluate their model, but they have to take certain liberties with the concepts involved. Specifically, Gómez *et al.* [2] map advertisements in their model to the certainty of opinions in ART. However, advertisements, which mean the recommendations from the providers themselves, cannot perfectly be translated into certainty in ART, which means the confidence of opinion providers about their opinions.

Harbers *et al.*'s trust model uses maximum entropy inference and minimum relative entropy to find and update the most probable probability distribution [3]. They point out that to evaluate their model in ART, they have to consider many variables that are not covered in their model. Besides, some features of their model cannot be verified in ART. They also observe that the performance of the trust model in ART depends too much on the opponents.

We face a similar challenge in evaluating our model using ART. First, in our model, trust values are built from either probability and certainty, or numbers of positive and negative outcomes. Unfortunately, agents can exchange only a probability in ART. Second, the goodness of a trust model according to ART depends on the agent's bank balance, which involves many variables not covered in our model. For example, in ART, an agent may say the painting is worth 1,000, while the actual value of the painting is 1,010. It is difficult to determine if this is a good estimate. Lastly, one of our main contributions is that our model provides aggregation and concatenation operators for agents to collect trust from multiple sources in a referral network. ART does not support agents providing referrals. Therefore, in order to evaluate our model, we design our own trust evaluation testbed.

3. SYSTEM ARCHITECTURE

We design a trust evaluation testbed where agents estimate the trustworthiness of other agents without preexisting knowledge. We define four kinds of agents:

Clients estimate the trust value of a provider with whom they have had no prior direct experience. A client asks witnesses who have direct experience with the provider. Clients find the witnesses via the recommenders who are their *neighbors*. Each client may also estimate trust placed in witnesses and recommenders.

Recommenders are agents who can provide referrals to witnesses or to other recommenders. Each recommender has other recommenders as its neighbors, and stores trust values of these neighbors. When a recommender receives a referral request from a client, it may (a) refer to witnesses, (b) refer to other recommenders, or (c) not provide any information. Each referral carries a trust value (a probability and certainty pair).

Witnesses are agents who have direct experience with providers. Each witness knows the sole provider in the system and calculates the expertise of the provider based on its past experience. The expertise is represented as a trust value.

(Service) Providers are agents with whom a client transacts. Each provider has a fixed expertise p. When a client has a transaction with the provider, the probability of a good outcome is p. The goal of the experiment is to show how accurately the client can estimate the trust value of the provider. For simplicity, we consider exactly one provider.

4. EVIDENCE-BASED TRUST MODEL

This paper extends W&S's model [5, 6] based on a PCDF [4]. We provide a mechanism to update the trust values of agents, based on the referrals they provide (Section 5).

In this experiment, an agent A estimates the trustworthiness of agent B by maintaining trust values of B. The trust value of B comes from both agent A's direct experience with agent B, and any referrals provided by a third-party, say, agent C. We model trust values in both the evidence space and the belief space. In evidence space, a trust value of an agent B is in the form $\langle r, s\rangle$, where $r \geq 0$ is the number of positive evidences with the agent B and $s \geq 0$ is the number of negative experiences with the agent B ($r + s \geq 0$). We define the probability $\alpha = \frac{r}{r+s}$, the expected value of the probability of a positive outcome. In the belief space, a trust value is modeled as a triple $\langle b, d, u\rangle$, where $b, d, u \geq 0$ and $b + d + u = 1$. The values of $\langle b, d, u\rangle$ can be interpreted as the weights of belief, disbelief, and uncertainty, respectively. The certainty $c = 1 - u$ represents the confidence of the probability. A bijective trust transformation from evidence to belief space is defined in [6].

W&S provide aggregation and concatenation operators (similar to Jøsang's consensus and recommendation operators [4]) that enable agent A to combine the trust values of B from different sources (e.g., direct experience with B and a referral provided by agent C).

The concatenation operator $\otimes$ is used when agent A collects a referral of agent B from agent C. The aggregation operator $\oplus$ is used when agent A combines trust values from different sources [5].

5. ADAPTIVE TRUST MODEL

In a referral network, agents evaluate trust values of the referrers by comparing the referrals they receive with their actual experience. For example, client C receives a referral of provider S from referrer R. Let C's current trust value of R be $M_R = \langle r_R, s_R\rangle$, and the trust value of S referred by R be $M_S = \langle r_S, s_S\rangle$. Suppose C's independent trust value of S is $M = \langle r, s\rangle$. C can update the trust value M_R of R, based on M_S and M. C's updated trust value $M'_R = \langle r'_R, s'_R\rangle$ of R can be defined as

$$q = \frac{\alpha_S^r(1-\alpha_S)^s}{\alpha^r(1-\alpha)^s} \quad (1)$$

$$r'_R = c_S q + (1-\beta) r_R \quad (2)$$

$$s'_R = c_S(1-q) + (1-\beta) s_R \quad (3)$$

The key intuition behind the above trust update rule is that C updates the trust value of referrer R by how close R's referred trust is to C's actual experience, discounted by R's certainty. Given the trust value $\langle r, s\rangle$, the PCDF is defined as $x^r(1-x)^s$. The maximum of the distribution occurs when x is equal to $\alpha = \frac{r}{r+s}$. Besides, if R's referred trust $\langle r_S, s_S\rangle$ is consistent with C's actual experience,

$\alpha_S = \frac{r_S}{r_S+s_S}$ should be $\alpha = \frac{r}{r+s}$. Then we can define the accuracy of the referred trust as the ratio q of the value of the probability α_S of the referred trust in the PCDF to the value of the probability α of the actual trust in the PCDF. We update the trust value $\langle r_R, s_R \rangle$ by concatenating the old $\langle r_R, s_R \rangle$ and $\langle c_S q, c_S(1-q) \rangle$. Note that the $\langle c_S q, c_S(1-q) \rangle$ is discounted by certainty c_S, because the referrers should be penalized (or rewarded) less if they provide referrals with low certainty.

The previous experience $\langle r_R, s_R \rangle$ is discounted by a time discounting rate $\beta \in [0, 1]$. With a high discount rate, the influence of agents' past behavior becomes less important, such that the updated trust values can reflect more on the current behavior. However, with a low discount rate, the updated trust values reflect the overall behavior of the target. Section 6.3 compares different discount rates, showing how discounting rate affects the agents' ability to deal with agents that change behavior over time.

6. EXPERIMENTAL RESULTS

In the experiment, each agent has two neighbors. For simplicity, there is only one client in this experiment. We compare the trust value of the provider estimated by clients with different trust models. In this experiment, each recommender can have two neighbors, and at most one witness.

At each simulation cycle, clients can ask their neighbors for recommendations. After receiving the recommendation requests from client agents, recommenders may refer to witnesses or other recommenders, or not reply at all. Clients can then build a referral tree by asking referred recommenders in advance, until the depth limit of the referral tree is reached or they find witnesses. At the end of each simulation cycle, clients' trust values of recommenders and witnesses are updated.

The experiment is initialized with one client, one provider, ten recommenders, and eight witnesses. Half of these recommenders are malicious, and half are honest. Four of the witnesses are honest, four are malicious. Honest agents provide referrals along with an accurate trust value, and malicious agents provide referrals along with an opposite trust value. For example, an honest recommender refers to an honest witness along with a trust value $\langle 19 \pm 1, 1 \pm 1 \rangle$, and a malicious witness refers to the good provider ($p = 1$) along with a trust value $\langle 1 \pm 1, 19 \pm 1 \rangle$. The neighbors of all agents are randomly assigned. Each recommender can refer to at most two other recommenders, or refer to one witness if the recommender knows one. The provider provides good service based on a fixed probability p. We define two providers: *damping* provider, which has $p = 1$ at the beginning and turns bad with $p = 0$ in the middle of the simulation, and *capricious* provider, which changes behavior alternating between $p = 1$ and $p = 0$ every two cycles. There are total 20 simulation cycles in one round. All results are the average of five rounds.

6.1 Comparing Update Models

To enable comparison, we extend Jøsang's model with a *linear* trust update method. The reason that the max-certainty trust update method cannot be applied in Jøsang's model is that Jøsang uses a different certainty definition from ours, which is the basis of our trust update method. The linear model represents trust as ours, but updates trust using the linear trust update, which defines q in equation 1 as $q = 1 - |\alpha - \alpha_S|$.

We define three types of clients: *max-certainty*, *linear*, and *Jøsang*, as shown in Table 1. The max-certainty model is the model described in Section 4. The Jøsang model uses the trust representation described in [4].

Model	Trust representation	Trust update
Max-certainty	Wang & Singh [6, 5]	Max-certainty
Linear	Wang & Singh [6, 5]	Linear
Jøsang	Jøsang [4]	Linear

Table 1: Models compared for collected trust values

6.2 Capricious Service Provider

In our first simulation, we evaluate our approach by the accuracy of the trust value of the capricious provider, honest agents (recommenders and witnesses), and malicious agents. The provider is capricious, which changes behavior between good and bad every two simulation cycles. There are five honest and five malicious recommenders. Four witnesses are honest, and four are malicious. The discount rate β of the trust update is 0.

Figure 1 shows the average probability α of the referred trust value (i.e., not direct experience) of the capricious provider over five simulation rounds. The max-certainty model provides the best estimate, which reaches 0.51 ± 0.01 after three cycles. The max-certainty model performs better than the linear and Jøsang models. The accuracy of referred trust value depends on the trust representation, combination operators, and the trust update method. The max-certainty uses different definition of the certainty and the trust update from the Jøsang model, and different definition of the trust update method from the linear model. Table 2 shows the last, the mean, and the standard deviation of the referred probability α, showing that the certainty definition and the max-certainty trust update method provide better trust estimation of the capricious agent.

Model	Final	Mean	Deviation
Actual	0.50	0.48	0.12
Max-certainty	0.52	0.51	0.01
Linear	0.56	0.54	0.02
Jøsang	0.65	0.76	0.11

Table 2: Comparing the probability α of the referred trust value of the capricious agent with three trust models

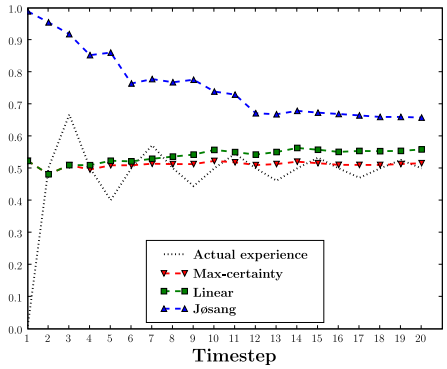

Figure 1: Comparing the probability α of the referred trust value of the capricious provider

6.3 Damping Provider

Damping agents are those who behave in a trustworthy manner at the beginning, and then turn to be malicious after their reputation has been built-up. One the most common ways of dealing with damping agents is to discount the trust value by time. In our max-certainty trust update method described in Section 5, β is the discount rate, which makes the update rule weigh past behavior less and current behavior more.

To show that our trust model has this property in dealing with damping agents, we construct a scenario where four of the eight recommenders are those who are honest during the first half of simulation, but turn to be malicious during the second half. These agents are damping recommenders. The rest four of the recommenders are always honest. Besides, we have four honest witnesses and a provider always providing good service ($p = 1.0$). We use different discount rates, 0.0, 0.4, 0.8, to show our model's flexibility of adjusting how fast reputation is destroyed.

Figure 2 shows the probability $\alpha = \frac{r}{r+s}$ and the certainty c of the trust value of a damping recommender with discount rate of $\beta = 0.0, 0.4$, and 0.8. With the discount rate $\beta = 0$, the probability α remains high while the certainty c grows slowly, as our client collects more positive evidence. But the probability α decreases at the middle of the simulation when the damping agent turns malicious. Note that the certainty c increases during the first half of the simulation as our client receives more evidence. When the damping recommender turns malicious, the certainty c drops, although the total evidence increases. This is because the evidence shows conflicts of the damping agent's behavior, which means the behavior of damping recommender is unstable. During the later half of the simulation, the certainty c grows slowly as our client collects more negative evidence of the damping recommender. At the end of the simulation, the probability α converges to 0.5, since half of the evidence is positive and the other half is negative.

When the discount rate β is 0.4, our trust model behaves similar to the result with discount rate $\beta = 0.0$, except the probability α drops faster when the damping recommender changes side. Moreover, the α converges to much lower than 0.5, because discount rate $\beta = 0.4$ makes the trust value reflect recent negative behavior more than previous positive behavior. On the other hand, when the discount rate β is 0.8, our trust model focuses on recent behavior even more, leading the probability α to drop even faster than discount rate $\beta = 0.4$. The certainty c, in this case, remains low, since we discount old evidence, and keep only new evidence.

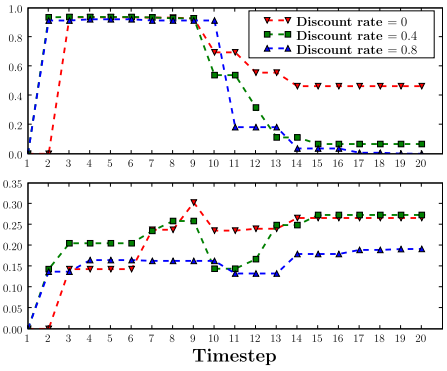

Figure 2: Comparing different discount rates β. The upper and lower plots show the probability α and certainty c of the referred trust value of a damping recommender, respectively.

6.4 Damping Provider with Different Models

Now we compare the three trust models with the same discount rate $\beta = 0.4$. Figure 3 shows that both the max-certainty and linear model provide good trust estimation of the damping provider. The difference of these two models is related to the trust update method. Although the linear trust update method helps the linear model have better correlation with the actual experience, the max-certainty trust update method provides more reasonable contrast of trust values between honest and malicious agents.

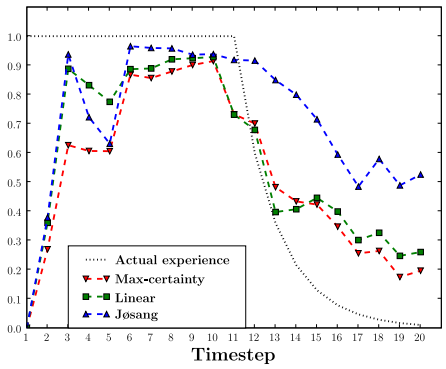

Figure 3: The averaged probability α and certainty c of trust value of a honest recommender

7. DISCUSSION

This paper presents an adaptive approach for probabilistic trust modeling for multiagent systems. Our model provides a probability-certainty representation of trust, trust collecting operators in referral networks, and the max-certainty trust update method for updating referrers. We design a simulation testbed to examine our model. Our results show that our model can estimate trust of agents well, even if those agents change behavior during the simulation. We also show that, by adjusting the discount rate, our approach can reflect the change of agents' behavior quickly.

However, our trust model does not show too much experimental advantage dealing with conflicts, even though it can handle conflicts theoretically. In future work, we plan to simplify the experiments, reducing the number of variables, for example, reducing the length of the referral chains in referral networks. Besides, we will seek a closed form for the optimal discount rate to facilitate applying our model in new domains.

8. REFERENCES

[1] K. K. Fullam, T. B. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. S. Barber, J. S. Rosenschein, L. Vercouter, and M. Voss. A specification of the agent reputation and trust (ART) testbed: experimentation and competition for trust in agent societies. In *Proc. 4th International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 512–518, 2005.

[2] M. Gómez, J. Carbó, and C. B. Earle. Honesty and trust revisited: the advantages of being neutral about other's cognitive models. *Autonomous Agents and Multi-Agent Systems*, 15(3):313–335, December 2007.

[3] M. Harbers, R. Verbrugge, C. Sierra, and J. Debenham. The examination of an information-based approach to trust. In *MALLOW Workshop on Coordination, Organization, Institutions and Norms in agent systems (COIN)*, September 2007.

[4] A. Jøsang. A subjective metric of authentication. In *Proc. 5th European Symposium on Research in Computer Security (ESORICS)*, 1998.

[5] Y. Wang and M. P. Singh. Trust representation and aggregation in a distributed agent system. In *Proc. 21st National Conference on Artificial Intelligence (AAAI)*, pp. 1425–1430, 2006.

[6] Y. Wang and M. P. Singh. Formal trust model for multiagent systems. In *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1551–1556, 2007.