A Probabilistic Model for Trust and Reputation

George Vogiatzis
Computer Science
Aston University
Birmingham, UK
g.vogiatzis@aston.ac.uk

Ian MacGillivray
Aston University
Birmingham, UK

Maria Chli
Computer Science
Aston University
Birmingham, UK
m.chli@aston.ac.uk

Abstract

This paper concerns the problem of agent trust in an electronic market place. We maintain that agent trust involves making decisions under uncertainty and therefore the phenomenon should be modelled probabilistically. We therefore propose a probabilistic framework that models agent interactions as a Hidden Markov Model (HMM). The observations of the HMM are the interaction outcomes and the hidden state is the underlying probability of a good outcome. The task of deciding whether to interact with another agent reduces to probabilistic inference of the current state of that agent given all previous interaction outcomes. The model is extended to include a probabilistic reputation system which involves agents gathering opinions about other agents and fusing them with their own beliefs. Our system is fully probabilistic and hence delivers the following improvements with respect to previous work: (a) the model assumptions are faithfully translated into algorithms; our system is optimal under those assumptions. (b) It can account for agents whose behaviour is not static with time (c) it can estimate the rate with which an agent's behaviour changes. The system is shown to significantly outperform previous state-of-the-art methods in several numerical experiments.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent Systems; I.6.5 [**Simulation and Modelling**]: Model Development, Modelling methodologies

General Terms

Algorithms

Keywords

Trust, reliability and reputation

1. INTRODUCTION

There has been a recent trend in computing towards increasingly larger, open and distributed systems. These systems are used by numerous human and software agents which act in a self-interested manner [16] and which often need to take decisions under uncertainty. Examples of such systems include, but are not limited to, electronic marketplaces, distributed file and information sharing systems, social and professional networking sites.

Acceptance of these systems depends upon average users achieving a useful, fair service of consistent quality. These aspirations are exceptionally hard to achieve given the difficulty of enforcing centralised control measures in distributed computing systems. Self-interested agents are likely to vary their behaviour according to their goals, even if this is at the expense of other users. Explicit penalties to any agents found to be compromising the system or behaving unfairly, are hard to impose and belie the ethos of an open, distributed paradigm. Instead, any agent who attempts to make profits through gaming or exploiting weaknesses in the workings of the system, should simply be flagged as an unsuitable choice and thereby lose opportunities for gain, fair or otherwise. A common approach to reaching this goal is the implementation of an automated trust and reputation system. Within this paper we consider trust, in line with the definition of [8], to be the belief an agent holds in the ability of its counterpart to deliver in the context of a specific transaction. This belief is based on the agent's previous interaction experience with its counterpart. Reputation is the collective opinion of others regarding an individual [16].

Researchers in the area of multi-agent societal issues have put forward a wide variety of representations for trust and reputation. The most prominent examples include different logic formalisms [1, 3, 11] as well as methods based on a mixture of probabilistic and empirical premises [3, 16]. Each of these systems is based on a set of assumptions and the translation of those assumptions into computer algorithms. Their performance is therefore related to both the validity of the assumptions as well as the algorithmic implementation. It is our postulation that only once a faithful representation of a set of assumptions has been made, can the effects and benefits of modifying these assumptions be studied under the scientific method. We further maintain that the only theoretically and practically sound language capable of representing uncertain reasoning is that of probability. Therefore the position we adopt in this paper is to derive a fully probabilistic model of trust from a simple set of assumptions. In doing so we eliminate any possibility of non-faithful translation of our assumptions into algorithms. Any shortcomings of the method presented can be solely attributed to faults within the assumptions themselves. The approach we propose as part of this work is, to the best of our knowledge, the

Cite as: A Probabilistic Model for Trust and Reputation, George Vogiatzis, Ian MacGillivray , Maria Chli, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 225-232

first one to maintain a fully probabilistic representation of trust and reputation while modelling agents with dynamic behaviour. A series of experimental evaluations demonstrate a significant performance improvement over state-of-the-art methods, thereby confirming the validity of our approach.

2. PREVIOUS WORK

A vast number of systems have been proposed to deal with the issue of computationally representing the intricate, semantically complex notion of trust (and its counterpart, reputation). One of the first works on the field was [9]. Dealing with deception was introduced in [7] while [15] considered the ability to combine sources of information.

Many of these systems have opposing views on how trust should be represented, stored and updated. Viewpoints vary between the Dempster-Shafer calculus [6], epistemic and doxastic logic [1], subjective logic[2] and fuzzy-logic [11].

Perhaps the most influential treatments of the problem of trust and reputation modelling are [3] and [16]. These models (like ours) are restricted to binary agent interaction outcomes. Trust is therefore the degree of belief an agent has into the outcome of an interaction being positive. Both models use the Beta distribution as a representation of the agent's posterior distribution in the interaction outcome, given all previous interaction experiences with another agent. The fundamental difference between these two treatments of trust is that [16] assumes static agents which does not always hold in real systems. On the other hand [3] relaxes this assumption by introducing a forgetting factor in the posterior update step. In [5, 10, 15] non-probabilistic models of trust and reputation are proposed which identify and attempt to cope with the problem of non-static agent behaviour.

A multi-dimensional generalization of [16] can be found in [14]. When dealing with a number of agents whose behaviour may be correlated, the authors of [14] propose modelling trust with a Dirichlet distribution instead of a set of independent Beta distributions. In this manner, correlations between agent interaction outcomes can be captured. As in [16], reputation is modelled by transmitting the sufficient statistics of these trust distributions between agents.

Another application of a trust system in the context of detecting malicious measurements in distributed sensor networks is provided in [12]. The Beta distributions are used in [12] to represent the belief of each sensor in the maliciousness of its parent sensor node.

Finally, the work described in [17] is very close in spirit to the work presented here. The authors consider the problem of selecting which agents to interact with based on previous experiences. Starting from a probabilistic model where posteriors on agent behaviour are updated based on experience, they formulate the decision theory problem which is then solved approximately. However, as in most other probabilistic treatments of trust, agents are assumed to have a constant behaviour. Here we show how changing agent behaviour can be taken into account within the probabilistic framework. Applying our methodology to the decision problem of [17] would be an interesting avenue for further research.

3. THE CONSTANT SERVICE PROVIDER

We first wish to establish a probabilistic representation of trust from first principles. We begin by considering only evidential trust; based on an agent's direct experiences alone. In this section we will assume that there is a single service provider with whom an agent has cause to interact. In our model an agent uses the service provider and can either be fully satisfied with the service or not. It is reasonably straightforward to extend these ideas to non-binary interactions but this is beyond the scope of the present paper.

In realistic scenarios it is reasonable to expect the service provider to change its behaviour with time corresponding with changes in management or policy decisions. We will however begin by considering the simpler case of the static service provider. This scenario is also the one considered in [16].

For a number of interactions, $1, \ldots, N$ the agent may record the discrete (good or bad) quality of the service (represented by 1 and 0 respectively) in variables $x_1, \ldots, x_N \in 0, 1$. Let H_π stand for the hypothesis that:

x_i is 1 with probability π for all i.

This may be expressed probabilistically as

$$p\left(x_i|H_\pi\right) = \pi^{x_i}\left(1-\pi\right)^{1-x_i}. \tag{1}$$

which is the standard Bernoulli trial distribution. We can also assume the conditional independence relation

$$p\left(x_i x_j|H_\pi\right) = p\left(x_i|H_\pi\right)p\left(x_j|H_\pi\right) \tag{2}$$

i.e., that outcomes of different interactions are probabilistically independent given the service provider's good service probability π. In the Bayesian Network formalism this is described by the graph shown in 1(a). An agent's trust for the service provider is a probability distribution for H_π given all evidence that the agent has seen. Initially, with no data, this distribution is a prior probability for H_π that expresses complete ignorance. This is represented by the uniform distribution; where all values for π are equally likely. Having observed $x_1, \ldots, x_t$, a prior probability for H_π may be inferred using Bayes' theorem[1] as:

$$\begin{aligned} p\left(H_\pi|x_1, \ldots, x_N\right) &= \frac{p\left(x_1, \ldots, x_N|H_\pi\right)p\left(H_\pi\right)}{p\left(x_1, \ldots, x_N\right)}. \\ &= C \times \prod_{i=1}^{N} p\left(x_i|H_\pi\right) \end{aligned} \tag{3}$$

where C is a constant that does not depend on H_π. Defining $G_t = x_1 + \ldots + x_t$:

$$p\left(H_\pi|x_1, \ldots, x_N\right) = C \times \pi^{G_t}\left(1-\pi\right)^{t-G_t} \tag{4}$$

and which may be estimated from the fact that:

$$\int_0^1 p\left(H_\pi|x_1, \ldots, x_N\right) d\pi = 1. \tag{5}$$

Equation (4) defines the posterior probability density for π after observing the outcome of $x_1, \ldots, x_N$. Whilst this model is useful for illustrating the basic strategy, it contains an inherent limitation; the assumption that a service provider's behaviour shall be constant throughout all future interactions. This assumption is rarely observed in practice for the reasons mentioned in Section 1. It can also be exploited by malicious service providers that provide good

[1] $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

services until they gain trust and subsequently provide poor services. The next section describes a more sophisticated model that incorporates change in π with time.

4. THE EVOLVING SERVICE PROVIDER

We change the model described in the previous section to incorporate a hidden, unobserved chain of good service probabilities for the same trader. To make inference possible we must further assume that the good trade probabilities do not change much between time instances. Instead of modelling these probabilities directly, we model a sequence of real numbers $y_1, \ldots, y_N \in \mathbb{R}$ that are converted into probabilities through the logistic function

$$\sigma(y) = \frac{1}{1+e^{-y}}. \tag{6}$$

The probability of obtaining a good service at time t is therefore taken to be $\sigma(y_t)$. As before, we define this through a binary Bernoulli trial model

$$p(x_t|y_t) = \sigma(y_t)^{x_t} (1-\sigma(y_t))^{1-x_t}. \tag{7}$$

We regard the numbers $y_1, \ldots, y_N$ as corresponding to the *quality* of the service provider at each time instant. If $y_y \gg 0$ then the service provider will almost certainly give a good service while if $y_t \ll 0$ the service will almost certainly be bad. Generally the aim of a trust system is to infer the quality of the service provider from the previous interaction outcomes $x_1, \ldots, x_N$. To make this inference possible we need to make some further assumptions on the service provider's quality variation through time. In particular we will assume that the service provider is slowly varying. This can be accomplished by assuming a Wiener process prior on the quality sequence as follows:

$$p(y_1 \cdots y_N) = p(y_1) \prod_{t=1}^{N} p(y_{t+1}|y_t) \tag{8}$$

where the conditional probabilities are given by

$$p(y_{t+1}|y_t) = N\left(y_{t+1}|y_t, \tau^2\right). \tag{9}$$

The parameter τ^2 is used to encode how much the quality of the service provider is considered to change from time t to $t+1$. The joint distribution of our full model then has the following form

$$p(x_1 \ldots x_N y_1 \ldots y_N) = p(y_1) \prod_{t=1}^{N-1} p(y_{t+1}|y_t) \prod_{t=1}^{N} p(x_t|y_t) \tag{10}$$

where $p(y_1)$ is our prior on the initial quality of the service provider. In our implementation this was modelled as a very wide Gaussian (e.g. variance 10) centred on zero. The joint distribution of equation 10 can be seen as a continuous-state Hidden Markov Model [13] with observations $x_1 \ldots x_N$ and hidden states $y_1 \ldots y_N$. In the Bayesian Network formalism this would be represented with the graph of figure 1(b). It is important to note that this model includes the Constant Trader model (outlined in section 3) as a special case when we take $\tau = 0$. Our strategy is (a) to use a recursion formula to infer the distribution of y_{N+1} given past observations $x_1 \ldots x_N$ and (b) obtain a maximum-likelihood estimate of the variance parameter τ^2 using Expectation-Maximization [4].

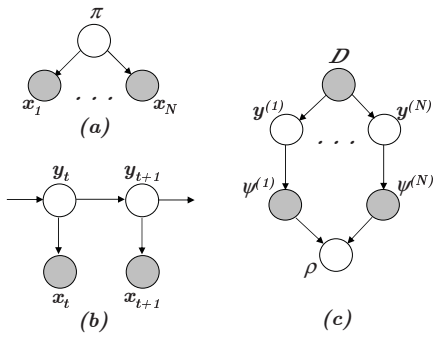

Figure 1: Bayesian Network representations of the three models presented in this paper: (a) the constant service provider (Section 3), (b) the evolving service provider (Section 4) and (c) the opinion provider model (Section 5). Arrows indicate conditional dependencies while shaded nodes represent observed variables.

4.1 Inference

The aim of a trust system is to infer the posterior distribution of the next quality level of the trader y_{N+1} given all past interaction outcomes $x_1 \ldots x_N$. From the expression of the joint distribution (10) one can derive the recursive formula

$$\begin{aligned} p(y_{t+1}|x_1 \ldots x_t) \quad \propto \quad & \int p(y_{t+1}|y_t) \\ & \times p(x_t|y_t) \\ & \times p(y_t|x_1 \ldots x_{t-1})\, dy_t. \end{aligned} \tag{11}$$

This expression links the posterior at time t, with the posterior at time $t-1$. By iterating this expression N times starting from the prior $p(y_1)$ we can obtain $p(y_{N+1}|x_1 \ldots x_N)$. From this posterior one can calculate the probability of the next trade being good given all previous trades by

$$\begin{aligned} p(x_{N+1}|x_1 \ldots x_N) \quad = \quad & \int_{-\infty}^{\infty} p(x_{N+1}|y_{N+1}) \\ & \times p(y_{N+1}|x_1 \ldots x_N)\, dy_{N+1}, \end{aligned} \tag{12}$$

as well as more complicated decision theoretic measures such as the expected loss if we decide to trade etc. The integrals of equations (11) and (12) do not have a closed form expression. However, because they are one-dimensional it is straightforward to use a variety of parametric approximations and integration formulas. Our approach is to model the distributions as histograms and approximate the integrals with summations.

4.2 Estimating service provider variance

The inference procedure described in the previous section assumes we have a good estimate of the variance parameter τ^2. In a realistic application of this model however, we would like to infer τ^2 from the data itself. In this paper we adopt a maximum-likelihood framework and we obtain a point estimate of τ^2 using the Expectation Maximization algorithm. In the particular context of the Hidden Markov Model, this is known as the forward-backward algorithm (Rabiner, 1989), or the Baum-Welch algorithm

(Baum, 1972). EM seeks to maximize the likelihood of observing the data $x_1 \dots x_N$ given the variance parameter τ^2

$$p\left(x_1 \dots x_N | \tau^2\right). \tag{13}$$

Define the quantities

$$\alpha_t\left(y_t\right) = p\left(x_1 \dots x_t y_t\right) \tag{14}$$

and

$$\beta_t\left(y_t\right) = p\left(x_{t+1} \dots x_N | y_t\right). \tag{15}$$

Once again we choose to model $\alpha_t\left(y_t\right)$ and $\beta_t\left(y_t\right)$ as histogram vectors. One can derive from (10) the following two recursion formulas. The forward recursion that provides $\alpha_t\left(y_t\right)$

$$\alpha_t\left(y_t\right) = p\left(x_t | y_t\right) \int \alpha_{t-1}\left(y_{t-1}\right) p\left(y_t | y_{t-1}\right) dy_{t-1} \tag{16}$$

and the backward recursion that provides $\beta_t\left(y_t\right)$

$$\beta_t\left(y_t\right) = \int p\left(x_{t+1} | y_{t+1}\right) \beta_{t+1}\left(y_{t+1}\right) p\left(y_{t+1} | y_t\right) dy_{t+1} \tag{17}$$

with initial values

$$\alpha_1\left(y_1\right) = p\left(x_1 | y_1\right) p\left(y_1\right) \tag{18}$$

and

$$\beta_N\left(y_N\right) = 1. \tag{19}$$

Now if we define

$$\xi_t\left(y_t, y_{t+1}\right) = \frac{1}{Q} \alpha_t\left(y_t\right) p\left(y_{t+1} | y_t\right) p\left(x_{t+1} | y_{t+1}\right) \beta_{t+1}\left(y_{t+1}\right) \tag{20}$$

with Q chosen so that

$$\iint \xi_t\left(y_t, y_{t+1}\right) dy_t dy_{t+1} = 1 \tag{21}$$

then one can write the EM update step as:

$$\tau^2_{new} = \frac{1}{N-1} \sum_{t=1}^{N-1} \iint \xi_t\left(y_t, y_{t+1}\right) \left(y_{t+1} - y_t\right)^2 dy_t dy_{t+1}. \tag{22}$$

Note that in equation (22) the distributions $\xi_t\left(y_t, y_{t+1}\right)$ have been computed with the old value for τ. Using (22) we obtain a new value τ_{new} which we then substitute back into the distributions α_t, β_t and ξ_t and repeat until convergence. Figure 2 shows a few iterations of this update step. With each update step the estimation becomes more and more accurate.

5. REPUTATION

In the absence of direct experience of a service provider that leads to direct *trust*, it is essential to be able to enquire about that service provider's *reputation*. In human societies this is commonly achieved by soliciting opinions about a particular service provider from others that may have had direct experience. In our model an agent will ask for opinions about a service provider from other agents and these opinions will be given as real numbers. This is in contrast to [3, 16] where agents transmit entire belief distributions to each other. From the point of view of the inquiring agent, an opinion ψ can be seen as a noisy measurement of the unknown state which is the service provider's quality y. Of course we have to take into account the possibility of agents not providing accurate opinions, either through lack of honesty or simple ignorance. For the purposes of our model, both these types of inaccuracy are bundled in the same category but it should be relatively straightforward to account for a more detailed opinion provider. To that end, we will consider the opinion ψ to be either:

- honest, in which case it is modelled by a Gaussian centred on y; this gives more likelihood to an honest opinion being close to the real value,
- or a dishonest one, in which case it can be uniformly distributed in an interval $[-w, w]$; this implies that a dishonest opinion is likely to be anything and has no connection with the real value.

With each opinion provider we will associate an *honesty ratio* ρ which is the likelihood that any one of his/her opinions about any service provider is honest. All of these assumptions can be summarized in the following mixture distribution

$$p\left(\psi | y, \rho\right) = \rho N\left(\psi | y, s^2\right) + \left(1 - \rho\right) U\left(\psi | -w, w\right) \tag{23}$$

where s^2 is the variance of an honest opinion and U is the uniform distribution pdf. In an alternative formulation we could also be transmitting an uncertainty measure as part of the opinion, and derive s^2 from that. However that makes the analysis significantly more complicated. At any rate, the results we have obtained seem to validate the simple approach of only transmitting ψ and assuming s known and constant. To keep the notation uncluttered we focus on a scenario with just *one* agent that requests opinions from *one* opinion provider regarding a set of K service providers. The extension to a scenario with multiple agents and multiple opinion providers will be straightforward. Denote the unknown qualities of each service provider with $Y = \left[y^{(1)}, \dots, y^{(K)}\right]$ and the opinions obtained with $\Psi = \left[\psi^{(1)}, \dots, \psi^{(K)}\right]$. Let ρ denote the unknown honesty ratio of our opinion provider. All previous interaction outcomes between our agent and each service provider can be summarized in a variable D. We assume that by solving the inference problem of section 4 our agent now has a set of posterior distributions describing its current knowledge about the service providers. Denote these posteriors by $p\left(y^{(1)} | D\right) \dots p\left(y^{(K)} | D\right)$. The full joint distribution of our reputation model is

$$p\left(Y, \Psi, \rho\right) = p\left(\rho\right) \prod_{k=1}^{K} p\left(\psi^{(k)} | y^{(k)}, \rho\right) p\left(y^{(k)} | D\right), \tag{24}$$

where $p\left(\rho\right)$ is the agent's prior about the honesty ratio of the opinion provider. In our implementation this starts out as a uniform prior. The Bayesian Network graph corresponding to this model can be found in Figure 1 (c). Our first task is to infer the posterior of the opinion provider's honesty ratio ρ given the opinions Ψ. Integrating out the unknown service provider states Y we get

$$\int p\left(Y, \Psi, \rho\right) dY = p\left(\rho\right) \prod_{k=1}^{K} \int p\left(\psi^{(k)} | y^{(k)}, \rho\right) p\left(y^{(k)} | D\right) dy^{(k)}. \tag{25}$$

The integral of eq. (25) becomes

$$\int p\left(\psi^{(k)}|y^{(k)},\rho\right)p\left(y^{(k)}|D\right)dy^{(k)} = \rho G^{(k)} + (1-\rho)\,B \tag{26}$$

where

$$G^{(k)} = \int N\left(\psi^{(k)}|y^{(k)},s^2\right)p\left(y^{(k)}|D\right)dy^{(k)} \tag{27}$$

and

$$B = U\left(\psi^{(k)}|-w,w\right). \tag{28}$$

Our posterior estimate of the honesty ratio is therefore

$$p\left(\rho|\Psi\right) \propto p\left(\rho\right)\prod_{k=1}^{K}\left(\rho G^{(k)} + (1-\rho)\,B\right) \tag{29}$$

where all the quantities are known and the proportionality constant can be obtained from the fact that $p\left(\rho|\Psi\right)$ integrates to 1. The distribution $p\left(\rho|\Psi\right)$ describes the agent's estimate about the accuracy of the opinions coming from this opinion provider. Our approach is superficially similar to the one of [16] where all opinions of an opinion provider are analyzed against past experience to evaluate this provider's accuracy. The difference is that our approach is fully probabilistic, i.e. starting from a set of predefined priors and likelihoods and arriving at posteriors using Bayes' theorem. The accuracy computation of [16] on the other hand is based on a set of empirical heuristics (e.g. the formula for adjusting an opinion according to the opinion provider's accuracy estimate). Section (6) contains a detailed comparison between [16] and our system's treatment of reputation.

Our second task is to take into account the opinions gathered by updating our quality posteriors. In particular we are looking for the new posteriors $p\left(y^{(k)}|\Psi,D\right)$ for $k = 1\ldots K$. Inverting with Bayes' theorem we get

$$\begin{aligned} p\left(y^{(k)}|\Psi,D\right) &= p\left(\psi^{(k)}|\Psi-\left\{\psi^{(k)}\right\},y^{(k)},D\right) \\ &\quad\times p\left(y^{(k)}|\Psi-\left\{\psi^{(k)}\right\},D\right). \\ &= p\left(\psi^{(k)}|\Psi-\left\{\psi^{(k)}\right\},y^{(k)},D\right) \\ &\quad\times p\left(y^{(k)}|D\right) \end{aligned} \tag{30}$$

also using the probabilistic independence of $\Psi-\left\{\psi^{(k)}\right\}$ with $y^{(k)}$ given D. Now the first term in this product can be expanded with

$$\begin{aligned} & p\left(\psi^{(k)}|\Psi-\left\{\psi^{(k)}\right\},y^{(k)},D\right) \\ &= \int p\left(\psi^{(k)}|\Psi-\left\{\psi^{(k)}\right\},y^{(k)},D,\rho\right) \\ &\quad\times p\left(\rho|\Psi-\left\{\psi^{(k)}\right\},y^{(k)},D\right)d\rho \\ &= \int p\left(\psi^{(k)}|y^{(k)},\rho\right) \\ &\quad\times p\left(\rho|\Psi-\left\{\psi^{(k)}\right\}\right)d\rho \end{aligned} \tag{31}$$

where we have used some probabilistic independence relations that can be derived from the factorization of the joint distribution (24). The distribution $p\left(\rho|\Psi-\left\{\psi^{(k)}\right\}\right)$ is the posterior for ρ that we get if we omit one of the opinions. It is given by

$$p\left(\rho|\Psi-\left\{\psi^{(k)}\right\}\right) \propto p\left(\rho\right)\prod_{\substack{j=1\\ j\neq k}}^{K}\left(\rho G^{(k)} + (1-\rho)\,B\right). \tag{32}$$

We now have all the ingredients to compute the updated quality posterior from eq. (30).

To extend this model to multiple opinion providers we treat each opinion provider sequentially. Each time we use eq. (30) to update our quality posteriors $p\left(y^{(k)}|D,\Psi_1,\Psi_2\ldots\right)$ to take into account each set of opinions. With each opinion from a particular opinion provider, we also update the honesty ratio posterior for that opinion provider using equation (29). In this manner the model generalizes to the setting where an agent interchangeably receives (a) observations of actual interactions with service providers and (b) opinions from opinion providers. The following section presents a detailed evaluation of an implementation of our model against the top competitors [3, 16].

6. EVALUATION

This section presents a series of numerical experiments designed to evaluate aspects of our model in comparison with the two of the main state-of-the-art competitors [3, 16]. We have separated the evaluation into the two components of our model, trust and reputation.

6.1 Evaluating trust

Our first experiment sheds light onto the performance of the EM algorithm for learning from data the rate of change of the service provider's quality. We created a random sequence of service qualities $y_1\ldots y_N$ (where $N = 500$) for a single service provider according to the Wiener process prior of equation (8) with $\tau = 0.15$ (red dotted line in figures 2(b-d). From this service quality sequence we then randomly sampled a sequence of binary interaction outcomes $x_1,\ldots,x_N$ according to the likelihood distribution of equation (7). This sequence was then fed to the EM algorithm described in section 4.2. We then ran a sequence of 1000 EM updates and recorded the evolution of the variance estimate. The algorithm converged at a value of 0.140 within the first 250 iterations. The convergence of the variance estimate is shown in 2 (a). For each value of τ we can compute the marginal posterior for all time-steps t using

$$p\left(y_t|x_1\ldots x_N\right) \propto \alpha_t\left(y_t\right)\beta_t\left(y_t\right) \tag{33}$$

with distributions α_t and β_t defined as per equations (14) and (15). Figures 2 (b-d) show the mean of this marginal posterior for each time-step together with ± 2 standard deviations away from the mean. We note how the estimate of the quality curve gradually becomes concentrated around the ground truth (shown by the red dotted line of Figures 2 (b-d)).

The second experiment consists of comparing the sequential service provider quality estimation of section 4.1 against TRAVOS [16]. This system assumes service providers of constant quality throughout the lifetime of the system. We randomly sampled a sequence from a constant trader ($\tau = 0$) (which fully satisfies the assumptions of [16]) and compared

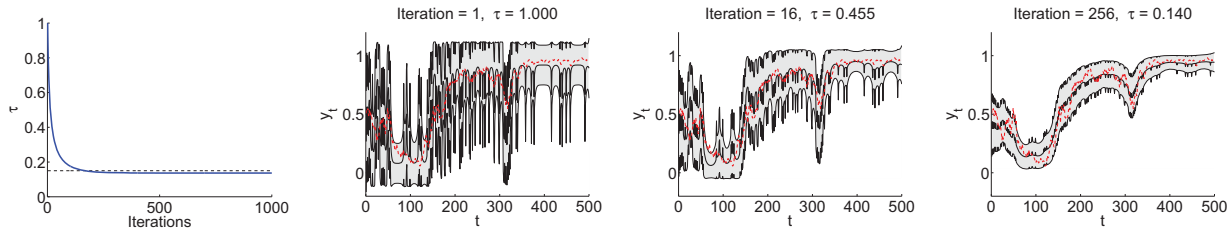

Figure 2: The EM estimation algorithm in practice. We estimated the rate of change parameter τ from a sequence of 500 interaction outcomes of a service provider. (a) shows the evolution of the τ estimate. The dotted line represents the actual value of τ used (0.15). Our estimate converges within 250 iterations to within 0.01 of the ground truth value. (b)-(d) show the posterior estimate of the service provider's quality for EM iterations 1, 16 and 256, respectively. Each graph shows the mean of the quality estimate along side two error margins (at ± 2 standard deviations away from the mean). We also superimpose with the red dotted line the ground truth evolution of the service provider's quality. Notice how the estimates of the quality curve get progressively better (more concentrated around the ground truth) as the EM algorithm converges.

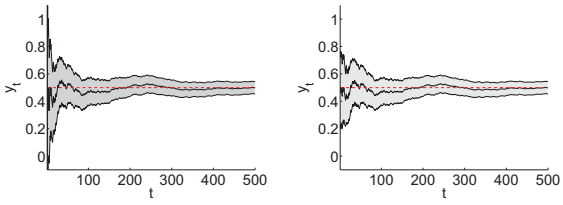

Figure 3: Our system (b) vs TRAVOS[16] (a) for a constant service provider. Our model contains the constant provider as a special case (for $\tau = 0$). When faced with such data, the EM algorithm estimates a τ value close to 0 and our inference results become identical to those of [16].

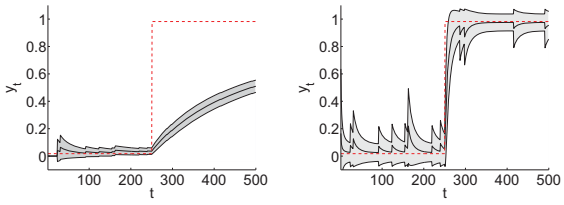

Figure 4: Our system (b) vs TRAVOS[16] (a) for a service provider that changes midway through the simulation. This case violates the assumptions of [16] (constant service providers) and hence the system cannot adapt to the change in quality of the service provider. Our model does not make any such assumption and is therefore able to track the step change in the quality.

our the performance of the two systems. As predicted in section 4 the constant quality model is a special case of the general evolving quality model. As a result, the EM algorithm was able to estimate that the quality variance for this sequence is very close to zero and hence the results obtained by our system are virtually identical to [16]. This is shown in figure 3 (a,b).

In the third experiment the service provider provided bad quality services for half the sequence, followed by good quality services for the remaining half. Since this falls outside the assumptions of [16], that system behaves quite poorly when estimating the quality of the service provider. At the same time, our system, after a period of uncertainty is able to track the variation in the quality. This is shown in figure 4.

Finally, in the forth experiment concerning the trust component of our model, we compare against the Beta Reputation System [3]. This system is theoretically able to cope with quality variations due to the introduction of a forgetting factor in the estimation stage. In fact that part is quite similar with our sequential inference of section 4.1 and the forgetting factor bears resemblance to our variance parameter τ. However, because in [3] some parts of the model do not strictly adhere to the probabilistic framework, the forgetting factor does not have a probabilistic interpretation. As a result, it is not possible to learn the correct value for the forgetting factor from the data as in our system. We generated data from a service provider whose quality varies in a sinusoidal fashion. Figure 5 (a,b) shows the difference between the two systems. For a particular value of the forgetting factor (found by trial and error) the curve obtained by [3] resembles the output of our method. However [3] does not provide any mechanism for estimating the forgetting factor from the data.

6.2 Evaluating reputation

In this section we compare our system with TRAVOS[16] with respect to the reputation part. The test setup comprises one agent that is faced against 100 service providers and 100 opinion providers. We assume that there has been a history of transactions between our agent and the 100 service providers. The agent's goal is to infer the quality of the 100 service providers using the opinions given by the 100 opinion providers. To focus the investigation on the reputation component we freeze all direct transactions between the agent and the service providers. Our agent has to rely only on his prior encounters and the set of opinions. Since the service providers are static in time for this experiment,

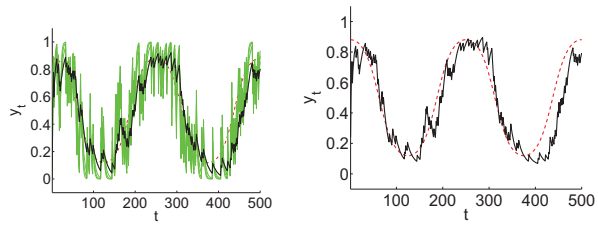

Figure 5: Our system (b) vs the Beta Reputation System [3] (a) for a service provider whose quality varies sinusoidally. The model of [3] can cope with varying service provider quality by introducing a forgetting factor parameter. Each parameter value produces a different estimate (green curves in (a)). By trial and error it is possible to find a parameter value that gives results (black curve in (a)) similar to those of our system (b). However, because the system is not fully probabilistic, there is no principled way to determine the optimal forgetting factor parameter. Our system on the other hand uses Expectation Maximization to obtain a maximum likelihood point-estimate of τ.

this fully satisfies the assumptions of [16].

The population of opinion providers in a real life situation may consist of a varied mix between honest, dishonest, well-informed and mis-informed members. We tested two different regimes, both of which gave similar results. In the first regime, we have two sub-populations. The first consists of entirely honest opinion providers whose opinions are accurate 90% of the time ($\rho = 0.9$), with a small 10% to account for honest transmission of erroneous information. The second sub-population consists of entirely dishonest opinion providers $\rho = 0$ who purposefully aim to misinform our agent. We explored various mixtures of honest-dishonest populations. In Figure 6 we show the results of our system compared with [16] for the $95\%-5\%$, $35\%-65\%$, $65\%-35\%$ and $5\%-95\%$ population mixes. In the graphs we have plotted the mean estimation error (MEE) [16] which is simply the mean absolute difference between the quality estimates of our agent and the actual quality estimates. The x-axis is the number of opinions gathered and varies from 0 (no opinions, just prior estimates) to 100 (all opinions plus the prior estimates). Of course both the [16] agent and the agent using our method used identical prior estimates. When 95% or 65% of opinion providers are honest, both systems are able to estimate the correct values quite quickly. However with fewer honest opinions the performance of [16] degrades while our system performs reasonable even with a very small percentage of honest opinion providers.

Under the second regime, the population consists of a homogeneous mixture of opinion providers, whose honesty ratio ρ (proportion of accurate opinions) is randomly sampled from a distribution centred around a particular value which we call *truthfulness* for the purposes of this experiment. Here we show the results for truthfulness values of 95%, 65%, 35% and 5%. Once again, as the overall population becomes more and more malicious, TRAVOS is unable to reach good final estimates while our system degrades far more gracefully. Ultimately both systems fail to converge to good estimates when the population is honest only in about 5% of the time.

The results seem to indicate quite clearly the validity of the fully probabilistic approach. The two systems start from almost identical assumptions but our system by staying strictly within the probabilistic framework is able to outperform the other system.

7. CONCLUSION

This paper presented a fully probabilistic model of trust and reputation. Under our framework, the task of deciding whether to interact with a service provider is one of inferring the current quality of that provider from all relevant information. Agents can exploit two sources of information: Firstly their prior experiences with an opinion provider and secondly, a set of opinions about that provider that are gathered from other agents. The system probabilistically fuses both types of information and obtains an estimate of the current quality of the service provider. Compared to two state-of-the-art systems, our approach shows a significant performance improvement, as evidenced in several experiments reported here. Furthermore, it brings the following benefits: (a) Our simple set of assumptions is accurately represented and incorporated in an algorithm. (b) Within the probabilistic framework it can accommodate agents whose quality changes with time and (c) it provides the means to estimate the rate of change of agent quality. We hope this investigation will serve as a proof of concept of the great benefits afforded by the fully probabilistic approach in the context of trust and reputation.

8. REFERENCES

[1] Herzig A., Lorini E., Hübner J. F., Ben-Naim J., Castelfranchi C., Demolombe R., Longin D., and Vercouter L. Prolegomena for a logic of trust and reputation. In Boella G., Pigozzi G., Singh M. P., and Verhagen H., editors, *NORMAS*, pages 143–157, 2008.

[2] Josang A. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–311, 2001.

[3] Josang A. and Ismail R. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, 2002.

[4] Dempster A. P., Laird N. M., and Rubin D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[5] Khosravifar B., Gomrokchi M., Bentahar J., and Thiran P. Maintenance-based trust for multi-agent systems. In *AAMAS '09: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 1017–1024, 2009.

[6] Yu B. and Singh M. P. Distributed reputation management for electronic commerce. *Computational Intelligence*, 18(4):535–549, 2002.

[7] Yu B. and Singh M. P. Detecting deception in reputation management. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 73–80, 2003.

[8] Gambetta D. *Trust: Making and breaking cooperative relations*, chapter Can we trust trust?, pages 213–237. Blackwell Publishers, 1988.

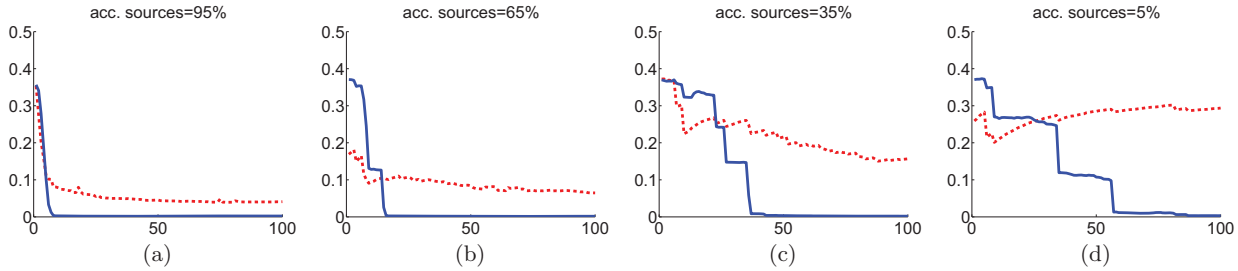

Figure 6: Reputation handling in our system vs TRAVOS [16]. The graphs show the mean estimation error of our system (solid blue) vs TRAVOS (dotted red). The x-axis corresponds to the number of opinions gathered. This experiment involves two sub-populations of opinion providers, an accurate and inaccurate one, mixed in various percentages. When most opinions are accurate both systems perform admirably. With fewer accurate opinions, the estimates our system obtains are superior.

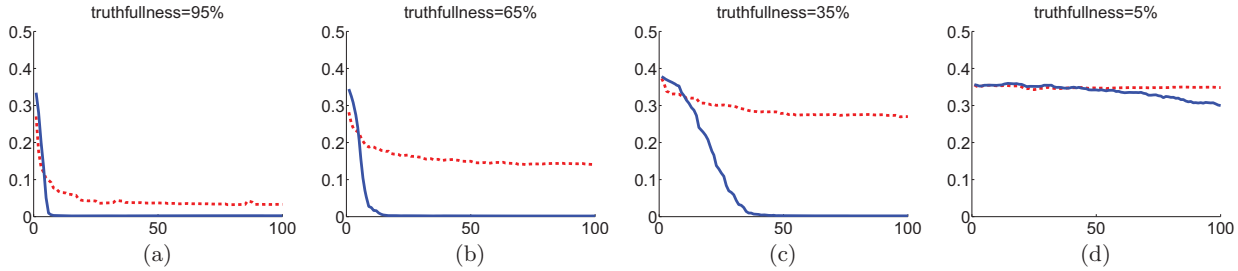

Figure 7: Reputation handling in our system vs TRAVOS [16]. This experiment concerns a homogeneous mixture of opinion providers who provide accurate opinions with a certain percentage. Similar results as in Figure 6 are observed.

[9] Zacharia G., Moukas A., and Maes P. Collaborative reputation mechanisms in electronic marketplaces. *Decision Support Systems*, 29(4):371–388, 2000.

[10] Pinyol I., Paolucci M., Sabater-Mir J., and Conte R. Beyond accuracy. reputation for partner selection with lies and retaliation. In *AAMAS '07: Proceedings of The 6th International Conference on Autonomous Agents and Multiagent Systems*, 2007.

[11] Rehak M., Pechoucek M., Grill M., and Bartos K. Trust-based classifier combination for network anomaly detection. In *Cooperative Information Agents XII*. Springer, 2008.

[12] Mistry O., Gursel A., and Sen S. Comparing trust mechanisms for monitoring aggregator nodes in sensor networks. In *AAMAS '09: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 985–992, 2009.

[13] Elliott R. J., Moore J. B., and Aggoun L. Hidden Markov Models: estimation and control, 1995.

[14] Reece S., Rogers A., Roberts S., and Jennings N. R. Rumours and reputation: Evaluating multi-dimensional trust within a decentralised reputation system. In *AAMAS '07: Proceedings of The 6th International Conference on Autonomous Agents and Multiagent Systems*, 2007.

[15] Huynh T. D., Jennings N. R., and Shadbolt N. R. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

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

[17] Teacy W. T. L., Chalkiadakis G., Rogers A., and Jennings N. R. Sequential decision making with untrustworthy service providers. In *AAMAS '08: Proceedings of The 7th International Conference on Autonomous Agents and Multiagent Systems*, pages 755–762, 2008.