A Map of Trust Between Trading Partners

John Debenham[1] and Carles Sierra[2]

[1] University of Technology, Sydney, Australia debenham@it.uts.edu.au
[2] Institut d'Investigacio en Intel.ligencia Artificial, Spanish Scientific Research Council, UAB
08193 Bellaterra, Catalonia, Spain sierra@iiia.csic.es

Abstract. A pair of 'trust maps' give a fine-grained view of an agent's accumulated, time-discounted belief that the enactment of commitments by another agent will be in-line with what was promised, and that the observed agent will act in a way that respects the confidentiality of previously passed information. The structure of these maps is defined in terms of a categorisation of utterances and the ontology. Various summary measures are then applied to these maps to give a succinct view of trust.

1 Introduction

The intuition here is that trust between two trading partners is derived by observing two types of behaviour. First, an agent exhibits trustworthy behaviour through the enactment of his commitments being in-line with what was promised, and second, it exhibits trustworthy behaviour by respecting the confidentiality of information passed 'in confidence'. Our agent observes both of these types of behaviour in another agent and represents each of them on a map. The structure of these two maps is defined in terms of both the type of behaviour observed and the ontology. The first 'map' of trust represents our agent's accumulated, time-discounted belief that the enactment of commitments will be in-line with what was promised. The second map represents our agent's accumulated, time-discounted belief that the observed agent will act in a way that fails to respect the confidentiality of previously passed information.

The only action that a software agent can perform is to send an utterance to another agent. So trust, and any other high-level description of behaviour, must be derived by observing this act of message passing. We use the term *private information* to refer to anything that one agent knows that is not known to the other. The intention of transmitting any utterance should be to convey some private information to the receiver — otherwise the communication is worthless. In this sense, trust is built through exchanging, and subsequently validating, private information [1]. Trust is seen in a broad sense as a measure of the strength of the relationship between two agents, where the *relationship* is the history of the utterances exchanged. To achieve this we categorise utterances as having a particular type and by reference to the ontology — this provides the structure for our map.

The literature on trust is enormous. The seminal paper [2] describe two approaches to trust: first, as a belief that another agent will do what it says it will, or will reciprocate for common good, and second, as constraints on the behaviour of agents to conform to trustworthy behaviour. The map described here is concerned with the first approach where trust is something that is learned and evolves, although this does not mean that we view the second as less important [3]. The map also includes reputation [4] that feeds into trust. [5] presents a comprehensive categorisation of trust research: policy-based, reputation-based, general *and* trust in information resources — for our trust maps, the estimating the integrity of information sources is fundamental. [6] presents an interesting taxonomy of trust models in terms of nine types of trust model. The scope described there fits well within the map described here with the possible exception of identity trust and security trust. [7] describes a powerful model that integrates interaction an role-based trust with witness and certified reputation that also relate closely to our model.

A key aspect of the behaviour of trading partners is the way in which they enact their commitments. The enactment of a contract is uncertain to some extent, and trust, precisely, is a measure of how uncertain the enactment of a contract is. Trust is therefore a *measure of expected deviations of behaviour* along a dimension determined by the type of the contract. A unified model of trust, reliability and reputation is described for a breed of agents that are grounded on information-based concepts [8]. This is in contrast with previous work that has focused on the similarity of offers [9, 10], game theory [11], or first-order logic [12].

We assume that a multiagent system $\{\alpha, \beta_1, \ldots, \beta_o, \xi, \theta_1, \ldots, \theta_t\}$, contains an agent α that interacts with negotiating agents, β_i, information providing agents, θ_j, and an *institutional agent*, ξ, that represents the institution where we assume the interactions happen [3]. Institutions provide a normative context that simplifies interaction. We understand agents as being built on top of two basic functionalities. First, a *proactive machinery*, that transforms *needs* into *goals* and these into *plans* composed of *actions*. Second, a reactive machinery, that uses the received messages to obtain a new world model by updating the probability distributions in it.

2 Ontology

In order to define a language to structure agent dialogues we need an ontology that includes a (minimum) repertoire of elements: a set of *concepts* (e.g. quantity, quality, material) organised in a is-a hierarchy (e.g. platypus is a mammal, Australian-dollar is a currency), and a set of relations over these concepts (e.g. price(beer,AUD)).[3] We model ontologies following an algebraic approach as:

An ontology is a tuple $O = (C, R, \leq, \sigma)$ where:

1. C is a finite set of concept symbols (including basic data types);
2. R is a finite set of relation symbols;
3. $\leq$ is a reflexive, transitive and anti-symmetric relation on C (a partial order)
4. $\sigma : R \to C^+$ is the function assigning to each relation symbol its arity

where $\leq$ is the traditional *is-a* hierarchy. To simplify computations in the computing of probability distributions we assume that there is a number of disjoint *is-a* trees covering different ontological spaces (e.g. a tree for types of fabric, a tree for shapes of clothing,

[3] Usually, a set of axioms defined over the concepts and relations is also required. We will omit this here.

and so on). R contains relations between the concepts in the hierarchy, this is needed to define 'objects' (e.g. deals) that are defined as a tuple of issues.

The semantic distance between concepts within an ontology depends on how far away they are in the structure defined by the $\leq$ relation. Semantic distance plays a fundamental role in strategies for information-based agency. How signed contracts, $Commit(\cdot)$, about objects in a particular semantic region, and their execution, $Done(\cdot)$, *affect* our decision making process about signing future contracts in nearby semantic regions is crucial to modelling the common sense that human beings apply in managing trading relationships. A measure [13] bases the *semantic similarity* between two concepts on the *path length* induced by $\leq$ (more distance in the $\leq$ graph means less semantic similarity), and the *depth* of the subsumer concept (common ancestor) in the shortest path between the two concepts (the deeper in the hierarchy, the closer the meaning of the concepts). Semantic similarity is then defined as:

$$\delta(c,c') = e^{-\kappa_1 l} \cdot \frac{e^{\kappa_2 h} - e^{-\kappa_2 h}}{e^{\kappa_2 h} + e^{-\kappa_2 h}}$$

where l is the length (i.e. number of hops) of the shortest path between the concepts c and c', h is the depth of the deepest concept subsuming both concepts, and κ_1 and κ_2 are parameters scaling the contributions of the shortest path length and the depth respectively.

3 Doing the 'right thing'

We now describe our first 'map' of the trust that represents our agent's accumulated, time-discounted belief that the enactment of commitments by another agent will be in-line with what was promised. This description is fairly convoluted. This sense of trust is built by continually observing the discrepancies, if any, between promise and enactment. So we describe:

1. How an utterance is represented in, and so changes, the world model.
2. How to estimate the 'reliability' of an utterance — this is required for the previous step.
3. How to measure the agent's accumulated evidence.
4. How to represent the measures of evidence on the map.

3.1 Updating the world model

α's world model consists of probability distributions that represent its uncertainty in the world's state. α is interested in the degree to which an utterance accurately describes what will subsequently be observed. All observations about the world are received as utterances from an all-truthful institution agent ξ. For example, if β communicates the goal "I am hungry" and the subsequent negotiation terminates with β purchasing a book from α (by ξ advising α that a certain amount of money has been credited to α's account) then α may conclude that the goal that β chose to satisfy was something other than hunger. So, α's world model contains probability distributions that represent its uncertain expectations of what will be observed on the basis of utterances received.

We represent the relationship between *utterance*, φ, and subsequent *observation*, φ', in the world model $\mathcal{M}^t$ by $\mathbb{P}^t(\varphi'|\varphi) \in \mathcal{M}^t$, where φ' and φ may be expressed in terms of ontological categories in the interest of computational feasibility. For example, if φ is "I will deliver a bucket of fish to you tomorrow" then the distribution $\mathbb{P}(\varphi'|\varphi)$ need not be over *all* possible things that β might do, but could be over ontological categories that summarise β's possible actions.

In the absence of in-coming utterances, the conditional probabilities, $\mathbb{P}^t(\varphi'|\varphi)$, tend to ignorance as represented by a *decay limit distribution* $\mathbb{D}(\varphi'|\varphi)$. α may have background knowledge concerning $\mathbb{D}(\varphi'|\varphi)$ as $t \to \infty$, otherwise α may assume that it has maximum entropy whilst being consistent with the data. In general, given a distribution, $\mathbb{P}^t(X_i)$, and a decay limit distribution $\mathbb{D}(X_i)$, $\mathbb{P}^t(X_i)$ decays by:

$$\mathbb{P}^{t+1}(X_i) = \Gamma_i(\mathbb{D}(X_i), \mathbb{P}^t(X_i)) \tag{1}$$

where Γ_i is the *decay function* for the X_i satisfying the property that $\lim_{t\to\infty} \mathbb{P}^t(X_i) = \mathbb{D}(X_i)$. For example, Γ_i could be linear: $\mathbb{P}^{t+1}(X_i) = (1-\varepsilon_i) \times \mathbb{D}(X_i) + \varepsilon_i \times \mathbb{P}^t(X_i)$, where $\varepsilon_i < 1$ is the decay rate for the i'th distribution. Either the decay function or the decay limit distribution could also be a function of time: Γ_i^t and $\mathbb{D}^t(X_i)$.

If α receives an utterance, μ, from β then: if α did not know μ already and had some way of accommodating μ then we would expect the integrity of $\mathcal{M}^t$ to increase. Suppose that α receives a message μ from agent β at time t. Suppose that this message states that something is so with probability z, and suppose that α attaches an epistemic belief $\mathbb{R}^t(\alpha,\beta,\mu)$ to μ — this probability reflects α's level of personal *caution* — a method for estimating $\mathbb{R}^t(\alpha,\beta,\mu)$ is given in Section 3.2. Each of α's active plans, s, contains constructors for a set of distributions in the world model $\{X_i\} \in \mathcal{M}^t$ together with associated *update functions*, $J_s(\cdot)$, such that $J_s^{X_i}(\mu)$ is a set of linear constraints on the posterior distribution for X_i. These update functions are the link between the communication language and the internal representation. Denote the prior distribution $\mathbb{P}^t(X_i)$ by p, and let $p_{(\mu)}$ be the distribution with minimum relative entropy[4] with respect to p: $p_{(\mu)} = \arg\min_r \sum_j r_j \log \frac{r_j}{p_j}$ that satisfies the constraints $J_s^{X_i}(\mu)$. Then let $q_{(\mu)}$ be the distribution:

$$q_{(\mu)} = \mathbb{R}^t(\alpha,\beta,\mu) \times p_{(\mu)} + (1 - \mathbb{R}^t(\alpha,\beta,\mu)) \times p \tag{2}$$

and to prevent uncertain observations from weakening the estimate let:

$$\mathbb{P}^t(X_{i(\mu)}) = \begin{cases} q_{(\mu)} & \text{if } q_{(\mu)} \text{ is more interesting than } p \\ p & \text{otherwise} \end{cases} \tag{3}$$

[4] Given a probability distribution q, the *minimum relative entropy distribution* $p = (p_1, \ldots, p_I)$ subject to a set of J linear constraints $g = \{g_j(p) = a_j \cdot p - c_j = 0\}$, $j = 1, \ldots, J$ (that must include the constraint $\sum_i p_i - 1 = 0$) is: $p = \arg\min_r \sum_j r_j \log \frac{r_j}{q_j}$. This may be calculated by introducing Lagrange multipliers λ: $L(p,\lambda) = \sum_j p_j \log \frac{p_j}{q_j} + \lambda \cdot g$. Minimising L, $\{\frac{\partial L}{\partial \lambda_j} = g_j(p) = 0\}$, $j = 1, \ldots, J$ is the set of given constraints g, and a solution to $\frac{\partial L}{\partial p_i} = 0, i = 1, \ldots, I$ leads eventually to p. Entropy-based inference is a form of Bayesian inference that is convenient when the data is sparse [14] and encapsulates common-sense reasoning [15].

A general measure of whether $q_{(\mu)}$ is more interesting than p is: $\mathbb{K}(q_{(\mu)}\|\mathbb{D}(X_i)) > \mathbb{K}(p\|\mathbb{D}(X_i))$, where $\mathbb{K}(x\|y) = \sum_j x_j \ln \frac{x_j}{y_j}$ is the Kullback-Leibler distance between two probability distributions x and y.

Finally merging Eqn. 3 and Eqn. 1 we obtain the method for updating a distribution X_i on receipt of a message μ:

$$\mathbb{P}^{t+1}(X_i) = \Gamma_i(\mathbb{D}(X_i), \mathbb{P}^t(X_{i(\mu)})) \tag{4}$$

This procedure deals with integrity decay, and with two probabilities: first, the probability z in the percept μ, and second the belief $\mathbb{R}^t(\alpha,\beta,\mu)$ that α attached to μ.

The interaction between agents α and β will involve β making contractual commitments and (perhaps implicitly) committing to the truth of information exchanged. No matter what these commitments are, α will be interested in any variation between β's commitment, φ, and what is actually observed (as advised by the institution agent ξ), as the enactment, φ'. We denote the relationship between commitment and enactment, $\mathbb{P}^t(\text{Observe}(\varphi')|\text{Commit}(\varphi))$ simply as $\mathbb{P}^t(\varphi'|\varphi) \in \mathcal{M}^t$.

In the absence of in-coming messages the conditional probabilities, $\mathbb{P}^t(\varphi'|\varphi)$, should tend to ignorance as represented by the *decay limit distribution* and Eqn. 1. We now show how Eqn. 4 may be used to revise $\mathbb{P}^t(\varphi'|\varphi)$ as observations are made. Let the set of possible enactments be $\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_m\}$ with prior distribution $p = \mathbb{P}^t(\varphi'|\varphi)$. Suppose that message μ is received, we estimate the posterior $p_{(\mu)} = (p_{(\mu)i})_{i=1}^m = \mathbb{P}^{t+1}(\varphi'|\varphi)$.

First, if $\mu = (\varphi_k, \varphi)$ is observed then α may use this observation to estimate $p_{(\varphi_k)k}$ as some value d at time $t+1$. We estimate the distribution $p_{(\varphi_k)}$ by applying the principle of minimum relative entropy as in Eqn. 4 with prior p, and the posterior $p_{(\varphi_k)} = (p_{(\varphi_k)j})_{j=1}^m$ satisfying the single constraint: $J^{(\varphi'|\varphi)}(\varphi_k) = \{p_{(\varphi_k)k} = d\}$.

Second, we consider the effect that the enactment ϕ' of another commitment ϕ, also by agent β, has on $p = \mathbb{P}^t(\varphi'|\varphi)$. Given the observation $\mu = (\phi', \phi)$, define the vector t as a linear function of semantic distance by:

$$t_i = \mathbb{P}^t(\varphi_i|\varphi) + (1 - \mid \delta(\phi', \phi) - \delta(\varphi_i, \varphi) \mid) \cdot \delta(\varphi', \phi)$$

for $i = 1, \ldots, m$. t is not a probability distribution. The multiplying factor $\delta(\varphi', \phi)$ limits the variation of probability to those formulae whose ontological context is not too far away from the observation. The posterior $p_{(\phi',\phi)}$ is defined to be the normalisation of t.

3.2 Estimating Reliability

$\mathbb{R}^t(\alpha,\beta,\mu)$ is an epistemic probability that takes account of α's personal caution. An empirical estimate of $\mathbb{R}^t(\alpha,\beta,\mu)$ may be obtained by measuring the 'difference' between commitment and enactment. Suppose that μ is received from agent β at time u and is verified by ξ as μ' at some later time t. Denote the prior $\mathbb{P}^u(X_i)$ by p. Let $p_{(\mu)}$ be the posterior minimum relative entropy distribution subject to the constraints $J_s^{X_i}(\mu)$, and let $p_{(\mu')}$ be that distribution subject to $J_s^{X_i}(\mu')$. We now estimate what $\mathbb{R}^u(\alpha,\beta,\mu)$ should have been in the light of knowing *now*, at time t, that μ should have been μ'.

The idea of Eqn. 2, is that $\mathbb{R}^t(\alpha,\beta,\mu)$ should be such that, *on average* across $\mathcal{M}^t$, $q_{(\mu)}$ will predict $p_{(\mu')}$ — no matter whether or not μ was used to update the distribution for X_i, as determined by the condition in Eqn. 3 at time u. The *observed belief* in μ and distribution X_i, $\mathbb{R}^t_{X_i}(\alpha,\beta,\mu)|\mu'$, on the basis of the verification of μ with μ', is the value of k that minimises the Kullback-Leibler distance:

$$\mathbb{R}^t_{X_i}(\alpha,\beta,\mu)|\mu' = \arg\min_k \mathbb{K}(k \cdot p_{(\mu)} + (1-k)\cdot p \parallel p_{(\mu')})$$

The predicted *information* in the enactment of μ with respect to X_i is:

$$\mathbb{I}^t_{X_i}(\alpha,\beta,\mu) = \mathbb{H}^t(X_i) - \mathbb{H}^t(X_{i(\mu)}) \tag{5}$$

that is the reduction in uncertainty in X_i where $\mathbb{H}(\cdot)$ is Shannon entropy. Eqn. 5 takes account of the value of $\mathbb{R}^t(\alpha,\beta,\mu)$.

If $\mathbf{X}(\mu)$ is the set of distributions that μ affects, then the *observed belief* in β's promises on the basis of the verification of μ with μ' is:

$$\mathbb{R}^t(\alpha,\beta,\mu)|\mu' = \frac{1}{|\mathbf{X}(\mu)|}\sum_i \mathbb{R}^t_{X_i}(\alpha,\beta,\mu)|\mu' \tag{6}$$

If $\mathbf{X}(\mu)$ are independent the predicted *information* in μ is:

$$\mathbb{I}^t(\alpha,\beta,\mu) = \sum_{X_i \in \mathbf{X}(\mu)} \mathbb{I}^t_{X_i}(\alpha,\beta,\mu) \tag{7}$$

Suppose α sends message μ to β where μ is α's private information, then assuming that β's reasoning apparatus mirrors α's, α can estimate $\mathbb{I}^t(\beta,\alpha,\mu)$. For each formula φ at time t when μ has been verified with μ', the *observed belief* that α has for agent β's promise φ is:

$$\mathbb{R}^{t+1}(\alpha,\beta,\varphi) = (1-\chi)\times\mathbb{R}^t(\alpha,\beta,\varphi) + \chi\times\mathbb{R}^t(\alpha,\beta,\mu)|\mu'\times\delta(\varphi,\mu)$$

where δ measures the semantic distance between two sections of the ontology as introduced in Section 2, and χ is the learning rate. Over time, α notes the context of the various μ received from β, and over the various combinations of utterance category, and position in the ontology, and aggregates the belief estimates accordingly. For example: "I believe John when he promises to deliver good cheese, but not when he is discussing the identity of his wine suppliers."

3.3 Measuring accumulated evidence

α's world model, $\mathcal{M}^t$, is a set of probability distributions. If at time t, α receives an utterance u that may alter this world model (as described in Section 3.1) then the (Shannon) *information* in u with respect to the distributions in $\mathcal{M}^t$ is: $\mathbb{I}(u) = \mathbb{H}(\mathcal{M}^t) - \mathbb{H}(\mathcal{M}^{t+1})$. Let $\mathcal{N}^t \subseteq \mathcal{M}^t$ be α's model of agent β. If β sends the utterance u to α then the *information* about β within u is: $\mathbb{H}(\mathcal{N}^t) - \mathbb{H}(\mathcal{N}^{t+1})$. We note that by defining information in terms of the change in uncertainty in $\mathcal{M}^t$ our measure is based on the way in which that update is performed that includes an estimate of the 'novelty' or 'interestingness' of utterances in Eqn 3.

3.4 Building the map

We give structure to the measurement of accumulated evidence using an *illocutionary framework* to categorise utterances, and an *ontology*. The illocutionary framework will depend on the nature of the interactions between the agents. The LOGIC framework for argumentative negotiation [16] is based on five categories: Legitimacy of the arguments, Options i.e. deals that are acceptable, Goals i.e. motivation for the negotiation, Independence i.e: outside options, and Commitments that the agent has including its assets. The LOGIC framework contains two models: first α's model of β's private information, and second, α's model of the private information that β has about α. Generally we assume that α has an illocutionary framework $\mathcal{F}$ and a categorising function $v : U \to \mathcal{P}(\mathcal{F})$ where U is the set of utterances. The power set, $\mathcal{P}(\mathcal{F})$, is required as some utterances belong to multiple categories. For example, in the LOGIC framework the utterance "I will not pay more for apples than the price that John charges" is categorised as both Option and Independence.

In [16] two central concepts are used to describe relationships and dialogues between a pair of agents. These are *intimacy* — degree of closeness, and *balance* — degree of fairness. Both of these concepts are summary measures of relationships and dialogues, and are expressed in the LOGIC framework as 5×2 matrices. A different and more general approach is now described. The intimacy of α's relationship with β_i, I^t_i, measures the amount that α knows about β_i's private information and is represented as real numeric values over $\mathcal{G} = \mathcal{F} \times \mathcal{O}$. Suppose α receives utterance u from β_i and that category $f \in v(u)$. For any concept $c \in \mathcal{O}$, define $\Delta(u,c) = \max_{c' \in u} \delta(c', c)$. Denote the value of I^t_i in position (f,c) by $I^t_{i(f,c)}$ then: $I^t_{i(f,c)} = \rho \times I^{t-1}_{i(f,c)} + (1-\rho) \times \mathbb{I}(u) \times \Delta(u,c)$ for any c, where ρ is the discount rate. The *balance* of α's relationship with β_i, B^t_i, is the element by element numeric difference of I^t_i and α's estimate of β_i's intimacy on α.

4 Not doing the 'wrong thing'

We now describe our second 'map' of the trust that represents our agent's accumulated, time-discounted belief that the observed agent will act in a way that fails to respect the confidentiality of previously passed information. Having built much of the machinery above, the description of the second map is simpler than the first.

[16] advocates the controlled revelation of information as a way of managing the intensity of relationships. Information that becomes public knowledge is worthless, and so respect of confidentiality is significant to maintaining the value of revealed private information. We have not yet described how to measure the extent to which one agent respects the confidentiality of another agent's information — that is, the strength of belief that another agent will respect the confidentially of my information: both by not passing it on, and by not using it so as to disadvantage me.

Consider the motivating example, α sells a case of apples to β at cost, and asks β to treat the deal in confidence. Moments later another agent β' asks α to quote on a case of apples — α might then reasonably increase his belief in the proposition that β had spoken to β'. Suppose further that α quotes β' a fair market price for the apples and that β' rejects the offer — α may decide to further increase this belief. Moments later β

offers to purchase another case of apples for the same cost. α may then believe that β may have struck a deal with β' over the possibility of a cheap case of apples.

This aspect of trust is the mirror image of trust that is built by an agent "doing the right thing" — here we measure the extent to which an agent does *not* do the wrong thing. As human experience shows, validating respect for confidentiality is a tricky business. In a sense this is the 'dark side' of trust. One proactive ploy is to start a false rumour and to observe how it spreads. The following reactive approach builds on the apples example above.

An agent will know when it passes confidential information to another, and it is reasonable to assume that the significance of the act of passing it on decreases in time. In this simple model we do not attempt to value the information passed as in Section 3.3. We simply note the amount of confidential information passed and observe any indications of a breach of confidence.

If α sends utterance u to β "in confidence", then u is categorised as f as described in Section 3.4. C_i^t measures the amount of confidential information that α passes to β_i in a similar way to the intimacy measure I_i^t described in Section 3.4: $C_{i(f,c)}^t = \rho \times C_{i(f,c)}^{t-1} + (1-\rho) \times \Delta(u,c)$, for any c where ρ is the discount rate; if no information is passed at time t then: $C_{i(f,c)}^t = \rho \times C_{i(f,c)}^{t-1}$. C_i^t represents the time-discounted amount of confidential information passed in the various categories.

α constructs a companion framework to C_i^t, L_i^t is as estimate of the amount of information leaked by β_i represented in $\mathcal{G}$. Having confided u in β_i, α designs update functions J_u^L for the L_i^t as described in Section 3.1. In the absence of evidence imported by the J_u^L functions, each value in L_i^t decays by: $L_{i(f,c)}^t = \xi \times L_{i(f,c)}^{t-1}$, where ξ is in $[0,1]$ and probably close to 1. The J_u^L functions scan every observable utterance, u', from each agent β' for evidence of leaking the information u, $J_u^L(u') = \mathbb{P}(\beta' \text{ knows } u \mid u' \text{ is observed})$. As previously: $L_{i(f,c)}^t = \xi \times L_{i(f,c)}^{t-1} + (1-\xi) \times J_u^L(u') \times \Delta(u,c)$ for any c.

This simple model estimates C_i^t the amount of confidential information passed, and L_i^t the amount of presumed leaked, confidential information represented over $\mathcal{G}$. The 'magic' is in the specification of the J_u^L functions. A more exotic model would estimate "who trusts who more than who with what information" — this is what we have elsewhere referred to as a *trust network* [17]. The feasibility of modelling a trust network depends substantially on how much detail each agent can observe in the interactions between other agents.

5 Summary Measures

[17] describes measures of: *trust* (in the execution of contracts), *honour* (validity of argumentation), and *reliability* (of information). The execution of contracts, soundness of argumentation and correctness of information are all represented as conditional probabilities $\mathbb{P}(\varphi'|\varphi)$ where φ is an expectation of what may occur, and φ' is the subsequent observation of what does occur.

These summary measures are all abstracted using the ontology; for example, "What is my trust of John for the supply of red wine?". These measures are also used to summarise the information in some of the categories in the illocutionary framework. For example, if these measures are used to summarise estimates $\mathbb{P}^t(\varphi'|\varphi)$ where φ is a deep motivation of β's (i.e. a Goal), or a summary of β's financial situation (i.e. a Commitment) then this contributes to a sense of trust at a deep social level.

The measures here generalise what are commonly called *trust*, *reliability* and *reputation* measures into a single computational framework. It they are applied to the execution of contracts they become trust measures, to the validation of information they become reliability measures, and to socially transmitted overall behaviour they become reputation measures.

Ideal enactments. Consider a distribution of enactments that represent α's "ideal" in the sense that it is the best that α could reasonably expect to happen. This distribution will be a function of α's *context* with β denoted by e, and is $\mathbb{P}^t_I(\varphi'|\varphi,e)$. Here we use relative entropy to measure the difference between this ideal distribution, $\mathbb{P}^t_I(\varphi'|\varphi,e)$, and the distribution of expected enactments, $\mathbb{P}^t(\varphi'|\varphi)$. That is:

$$M(\alpha,\beta,\varphi) = 1 - \sum_{\varphi'} \mathbb{P}^t_I(\varphi'|\varphi,e) \log \frac{\mathbb{P}^t_I(\varphi'|\varphi,e)}{\mathbb{P}^t(\varphi'|\varphi)} \tag{8}$$

where the "1" is an arbitrarily chosen constant being the maximum value that this measure may have.

Preferred enactments. Here we measure the extent to which the enactment φ' is preferable to the commitment φ. Given a predicate $\text{Prefer}(c_1,c_2,e)$ meaning that α prefers c_1 to c_2 in environment e. An evaluation of $\mathbb{P}^t(\text{Prefer}(c_1,c_2,e))$ may be defined using $\delta(\cdot)$ and the evaluation function $w(\cdot)$ — but we do not detail it here. Then if $\varphi \leq o$:

$$M(\alpha,\beta,\varphi) = \sum_{\varphi'} \mathbb{P}^t(\text{Prefer}(\varphi',\varphi,o))\mathbb{P}^t(\varphi' \mid \varphi)$$

Certainty in enactment. Here we measure the consistency in expected acceptable enactment of commitments, or "the lack of expected uncertainty in those possible enactments that are better than the commitment as specified". If $\varphi \leq o$ let: $\Phi_+(\varphi,o,\kappa) = \{\varphi' \mid \mathbb{P}^t(\text{Prefer}(\varphi',\varphi,o)) > \kappa\}$ for some constant κ, and:

$$M(\alpha,\beta,\varphi) = 1 + \frac{1}{B^*} \cdot \sum_{\varphi' \in \Phi_+(\varphi,o,\kappa)} \mathbb{P}^t_+(\varphi'|\varphi) \log \mathbb{P}^t_+(\varphi'|\varphi)$$

where $\mathbb{P}^t_+(\varphi'|\varphi)$ is the normalisation of $\mathbb{P}^t(\varphi'|\varphi)$ for $\varphi' \in \Phi_+(\varphi,o,\kappa)$,

$$B^* = \begin{cases} 1 & \text{if } |\Phi_+(\varphi,o,\kappa)| = 1 \\ \log|\Phi_+(\varphi,o,\kappa)| & \text{otherwise} \end{cases}$$

6 Conclusion

Trust is evaluated by applying summary measures to a rich model of interaction that is encapsulated in two maps. The first map gives a fine-grained view of an agent's accumulated, time-discounted belief that the enactment of commitments by another agent will be in-line with what was promised. The second map contains estimates of the accumulated, time-discounted belief that the observed agent will act in a way that fails to respect the confidentiality of previously passed information. The structure of these maps is defined in terms of a categorisation of utterances and the ontology. Three summary measures are described that may be used to give a succinct view of trust.

References

1. Reece, S., Rogers, A., Roberts, S., Jennings, N.R.: Rumours and reputation: Evaluating multi-dimensional trust within a decentralised reputation system. In: 6th International Joint Conference on Autonomous Agents and Multi-agent Systems AAMAS-07. (2007)
2. Ramchurn, S., Huynh, T., Jennings, N.: Trust in multi-agent systems. The Knowledge Engineering Review **19** (2004) 1–25
3. Arcos, J.L., Esteva, M., Noriega, P., Rodríguez, J.A., Sierra, C.: Environment engineering for multiagent systems. Journal on Engineering Applications of Artificial Intelligence **18** (2005)
4. Sabater, J., Sierra, C.: Review on computational trust and reputation models. Artificial Intelligence Review **24** (2005) 33–60
5. Artz, D., Gil, Y.: A survey of trust in computer science and the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web **5** (2007) 58–71
6. Viljanen, L.: Towards an ontology of trust. In Katsikas, S., Løpez, J., Pernum, G., eds.: Trust, Privacy and Security in Digital Business TrustBus 2005, Springer-Verlag (2005) 175–184
7. Huynh, T., Jennings, N., Shadbolt, N.: An integrated trust and reputation model for open multi-agent systems. Autonomous Agents and Multi-Agent Systems **13** (2006) 119–154
8. MacKay, D.: Information Theory, Inference and Learning Algorithms. Cambridge University Press (2003)
9. Jennings, N., Faratin, P., Lomuscio, A., Parsons, S., Sierra, C., Wooldridge, M.: Automated negotiation: Prospects, methods and challenges. International Journal of Group Decision and Negotiation **10** (2001) 199–215
10. Faratin, P., Sierra, C., Jennings, N.: Using similarity criteria to make issue trade-offs in automated negotiation. Journal of Artificial Intelligence **142** (2003) 205–237
11. Rosenschein, J.S., Zlotkin, G.: Rules of Encounter. The MIT Press, Cambridge, USA (1994)
12. Kraus, S.: Negotiation and cooperation in multi-agent environments. Artificial Intelligence **94** (1997) 79–97
13. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering **15** (2003) 871 – 882
14. Cheeseman, P., Stutz, J.: On The Relationship between Bayesian and Maximum Entropy Inference. In: Bayesian Inference and Maximum Entropy Methods in Science and Engineering. American Institute of Physics, Melville, NY, USA (2004) 445 – 461
15. Paris, J.: Common sense and maximum entropy. Synthese **117** (1999) 75 – 93
16. Sierra, C., Debenham, J.: The LOGIC Negotiation Model. In: Proceedings Sixth International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2007, Honolulu, Hawai'i (2007)
17. Sierra, C., Debenham, J.: Trust and honour in information-based agency. In Stone, P., Weiss, G., eds.: Proceedings Fifth International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2006, Hakodate, Japan, ACM Press, New York (2006) 1225 – 1232