OpenKnowledge

FP6-027253

Trust and Reputation

Sindhu Joseph[1], Carles Sierra[1], and Fausto Giunchiglia[2]

[1] Artificial Intelligence Research Institute, IIIA-CSIC, Spain
[2] Dept of Information and Communication Technology, University of Trento, Italy

Report Version: final
Report Preparation Date: 30.1.2007
Classification: deliverable D4.7
Contract Start Date: 1.1.2006 Duration: 36 months
Project Co-ordinator: University of Edinburgh (David Robertson)
Partners: IIIA(CSIC) Barcelona
Vrije Universiteit Amsterdam
University of Edinburgh
KMI, Open University
University of Southampton
University of Trento

Trust models for open MAS

Sindhu Joseph, Carles Sierra
Institut d'Investigacio en Intel.ligencia Artificial
Spanish Scientific Research Council, UAB
08193 Bellaterra, Catalonia, Spain
sierra@iiia.csic.es

Fausto Giunchiglia
Department of Information and Communication Technology
University of Trento
fausto@dit.unitn.it

January 30, 2007

1 Introduction

What does it mean for a society of agents or an electronic community to be open? A possible answer is to be neutral with respect to the architecture of the system. Another one is that from the prospect of agent technology, we may have agents without an assigned goal and the goals assigned to agents are, in general, loose. This society is open to new agents either with no definite goal or with self motivated goals not exceedingly relevant to the society. In other words they assume the heterogeneity of the participating agents.This openness signifies that the architecture of the system is not driven by goals imposed to the collective of agents, but that general goals arise from the collective actions of agents. Also, the adjunction of new agents must be as easy as required. In other words, the architecture of the society can exhibit some dynamical features rather than being solely static or bounded. Such a society or community should have a basic ingredient to be sustainable, *a computational mechanism for trust.*

The scientific research in the area of computational mechanisms for trust and reputation is a recent discipline oriented to increase the reliability and performance of electronic communities. The new paradigm of the so called intelligent or autonomous agents and Multi-Agent Systems (MAS) together with the spectacular emergence of the information society technologies (specially reflected by the popularization of electronic commerce) are responsible for the increasing interest on trust and reputation mechanisms applied to electronic societies.

An important scenario, that needs particular mention is that of Negotiation. Negotiation is a fundamental concept in multi-agent systems because it enables (self-interested) agents to find agreements and partition resources effeciently and effectively [34, 20]. Recently, a number of automated negotiation mechanisms have been proposed ranging from centralised approaches, using mechanism design or auction theory [31, 14], to distributed approaches using bargaining protocols [28]. In all of these approaches, however, the essence of the agent interactions boils down to making commitments to (contracts with) one another to carry out particular tasks. A commitment in this sense is a pledge to abide by the conditions set out in the agreement [40]. These commitments may involve, for example, a pledge to pay a particular amount of money for a service or a pledge to deliver goods within a particular time frame. However, in open environments (e.g. the grid [15], the semantic web [39], ubiquitous computing [2], and e-commerce [29]) where agents are probably at their most useful, there is no guarantee that a contracted agent will actually fulfill its commitments (even if there are associated penalty clauses for reneging). This is because there are a number of factors that might entice it to renege on its commitments:

1. The agents often represent different (self-interested) stakeholders, each with its own aims and objectives. This means the most common design strategy for an agent is to maximise its expected individual utility [26]. This may, in turn, well involve breaking earlier commitments if more profitable opportunities present themselves.

2. Given the scale and distributed nature of the system, agents are unlikely to have complete information about their counterparts. Therefore, in taking up an offered service, a client may not know how effective and/or efficient the provider is in actually delivering that service. This means there could be a gap between the service as advertised and the service as it is actually delivered that the agent could exploit to gain benefit. In the extreme case, an agent may simply renege on the whole deal (i.e. not deliver any of the agreed service), while in other cases an agent may only *partly* fulfill its commitments (i.e. up to a certain degree rather than completely renege on them). For example, several seller agents might offer the same product but each may have different degrees of efficiency with respect to delivering on time. Therefore, in this case the later the delivery time, the lower the degree of fulfillment of the sellers' commitment to deliver the product (at the agreed time).

Given this background, it is clear that agents are faced with significant degrees of uncertainty about the efficiency and effectivness of their counterparts in enacting the terms of the contract they wish to negotiate. In general, the estimates of uncertainty that an agent has about its (possible) interaction partners can be captured through the overarching notion of trust.

In the OpenKnowledge context this is especially true, as the most acclaimed notion is its openness. The services are modeled as interaction protocols in an electronic institution which is executed using the light weight protocol language

(LCC). When an interaction model is defined, to execute the model, it is required to find agents that can enact roles in the model. Agents need to have capabilities to fulfill these roles, for the successful execution. Yet, as we have an open system, and as we assume heterogeneity of agents, it can happen that agents can lie about their abilities. As there is no other mechanism to verify such behavior, trust and reputation models help us evaluate the trustworthiness of an agent.

Another aspect that is relevant to open systems and in particular to OpenKnowledge is the trust that one places on the services, software components and in general on the interaction models that are found in the open community. The difference between trust on an agent and trust on a service may be conceptual. Trust on a software entity more or less can be mapped to a number of factors describing the quality of the entity. This can be for example, the correctness, efficiency of the design, optimization whether performed, etc. On a different perspective, the trust on a software entity can be entitled to the trust on the owner of that entity. This works very well if the open system associates an entity to its owner. This may be a reasonable assumption, as one of the predominant open systems, the internet does so. In the latter case, the computational mechanisms for trust and reputation on agents can be applied directly for software entities too. In the former case, each entity needs to have a QOS(quality of service) associated with it, which can be computed in various ways. A discussion of these methods is outside the scope of this paper. But just to complete our argument, one of the ways in getting this value is to use the value provided by the owner, in which case, the trust on the value can be translated to the trust on the owner. Thus to conclude, computational mechanisms for trust on agents can to a great extend take care of trust on software artifacts generated by them. The difference in treatment could only be in terms of how extensive a model is required. So from now on in the discussions, when trust on agents is mentioned we assume the same is applicable to trust on artifacts unless explicitly specified otherwise.

In the following sections we present a summary of the important works in the area of computational trust and reputation in the recent years along with a set of criteria for classification of such models. Later sections provide a detailed illustration of three trust models namely the CREDIT, ReGreT and Information Theory based model that have been developed by researchers of the IIIA and that provide the baseline for trust modelling within Open Knowledge. In deliverable D4.4 a discussion on how these models can support semantic matching is made.

2 Criteria for model classification

Although the study of computational trust and reputation models is quite recent, in the last few years a lot of different proposals have appeared. These models can be classified based on the following factors: Information sources considered (*direct experience, witness information, sociological information, prejudice etc*), visibility types (*subjective, objective*), granularity of the model (*single* *context, multi context*), agent types and reliability measures. The following elaborates this

2.1 Classification Dimensions

- As would be expected, the main information sources used by the trust and reputation models are direct experiences and information from third party agents (witness information). There are very few models that take into account other aspects to calculate trust and reputation values. These two sources of information are, with no doubt, the most relevant. Nonetheless, a good mechanism to increase the efficiency of actual trust and reputation models (and also to overcome the lack of condence in e-markets) is the introduction of sociological aspects as part of these models.

- The visibility of the trust and reputation of an individual can either be seen as global shared by all the observers or as subjective assessed particularly by each individual. In the first case, the trust/reputation value is calculated from the opinions of the individuals that in the past interacted with the individual being evaluated. This value is publicly available to all members of the community and updated each time a member issues a new evaluation of an individual. In the second case, each individual assigns a personalized trust/reputation value to each member of the community according to more personal elements like direct experiences, information gathered from witnesses, known relations between members of the community and so on.

- Likewise the granularity of the models could be context dependent (multi context) or otherwise (single context). A single-context trust/reputation model is designed to associate a single trust/reputation value per partner without taking into account the context. A multi-context model has the mechanisms to deal with several contexts at a time maintaining different trust/reputation values associated to these contexts for a single partner.

- The models could be classified according to the degree of sophistication in agent behavior. The simplest model assumes all agents to be honest at all times. A more sophisticated model assumes that agents can hide information though they never lie. Finally there are models with specific mechanism to take care of liars.

- As important as the trust/reputation value itself is to know how reliable is that value and the relevance it deserves in the final decision making process. Depending on the model, the elements that are considered to calculate the reliability measure are different. Among them there may be elements like the number of experiences, the reliability of witnesses, how old is the information used to build trust and reputation, and so on.

2.2 Contributions

In this section we summarise the most influential works in the field.

The trust model proposed by Marsh [25] is one of the earliest. The model only takes into account direct interaction. It differenciates three types of trust: Basic trust which models the general trusting disposition independently of who is the agent that is in front, General trust which is the trust that one agent has on another without taking into account any specic situation, and Situational trust which is the amount of trust that one agent has in another taking into account a specic situation. The utility of the situation, its importance and the General trust are the elements considered in order to calculate the Situational trust.

eBay [11], Amazon Auctions [3] and OnSale Exchange [27] are good examples of online marketplaces that use reputation mechanisms. eBay is one of the worlds largest online marketplace with a community of over 50 million registered users. Most items on eBay are sold through English auctions and the reputation mechanism used is based on the ratings that users perform after the completion of a transaction. The user can give three possible values: positive(1), negative(-1) or neutral(0). The reputation value is computed as the sum of those ratings over the last six months. Similarly, Amazon Auctions and OnSale Exchange use also a mean (in this case of all ratings) to assign a reputation value. All these models use reputation as a global property. Though simple, this has definitely contributed to the success of e-markets.

Sporas and Histos [42] are evolved versions of the online reputation models discussed above. Histos includes witness information in the calculation of trust values. They also incorporate a reliability measure of the trust values produced.

The trust model proposed by Schillo et al. [36] is intended for scenarios where the result of an interaction between two agents (from the point of view of trust) is a boolean impression. The model is based on probability theory and proposes a Prisoners dilemma set of games to calculate the trust value of a partner. This model also introduces the *TrustNet* concept. It is used by each individual agent to collect witness information in a systematic manner. Different from models discussed so far, this model takes care of information hiding agent behavior. The author provides no information however concerning how to combine direct experience with witness information. It also does not take care of contexts.

This trust model Abdul-Rahman and Hailes [1] uses four degrees of belief to typify agent trustworthiness: vt (very trustworthy), t (trustworthy), u (untrustworthy) and vu (very untrustworthy). For each partner and context, the agent maintains a tuple with the number of past experiences in each category. Then, from the point of view of direct interaction, the trust on a partner in a given context is equal to the degree that corresponds to the maximum value in the tuple. Contrary to other trust models where witness information is merged with direct information to obtain the trust on the specic subject, this model is intended to evaluate only the trust on the information given by witnesses. Direct experiences are used to compare the point of view of these witnesses with the direct perception of the agent and then be able to adjust the information coming from them accordingly.

The trust model proposed by Esfandiari and Chandrasekharan [12], combine direct observation with interaction. Bayesian learning is used to update direct observation trust values. There are two main protocols of interaction, the exploratory protocol where the agent asks the others about known things to evaluate their degree of trust and the query protocol where the agent asks for advice from trusted agents. This model takes care of multi contexts and proposes a trust acquisition mechanism called institutional trust, the idea being to exploit the structure in the environment to determine trust values. However as in many models, there is no information provided to combine the different trust acquisition mechanisms.

In Sen and Sajjas' [38] reputation model, both interaction and observation types of direct experiences are considered. The focus of this work is on how agents use word of mouth information to assign trust and reputation values. Reinforcement learning is used to update the reputation value. A mechanism to take care of liars though it is assumed that liars lie consistently.

The main characteristic of the model by Carbo [6] is the use of fuzzy sets to represent reputation values. They combine old and new fuzzy values with a weighted aggregation and is termed as remembrance or memory. This allows to give more importance to the latest interaction. The notion of reliability of the reputation value is modeled through the fuzzy sets themselves. Recommendations from other agents are aggregated directly with the direct experiences.

The main idea behind the reputation model presented by Carter et al. [7] is that the reputation of an agent is based on the degree of fulfillment of roles ascribed to it by the society. As these roles are local to a society, it is impossible to universalize the calculation of reputation. The users overall reputation is calculated as a weighted aggregation of the degree of fulfillment of each role where the weights are entirely dependent on the specific society.

The trust model proposed by Castelfranchi and Falcone [8] is a clear example of a cognitive trust model. The basis of their model is the strong relation between trust and delegation. That is the decision that takes an agent x to delegate a task to agent y is based on a specic set of beliefs and goals and this mental state is termed as trust. They define the set of beliefs to build up a state of trust as competence, willingness, dependence, persistence, etc.

Following sections present three different models for computational trust, in the order of increasing sophistication. The context of most of the discussion that follows is negotiation dialogues. Yet negotiation dialogues need not be viewed in its narrow setting of electronic commerce, but can be viewed in a broader spectrum, where services and information are negotiated. In the context of OpenKnowledge negotiation dialogues are for services. Services advertised by peers as interaction models. Quality of service(QOS), utility of information, time constraints, and availability are some of the basis on which negotiation takes place.

3 ReGreT

Up to now, the computational models of trust and reputation have been considering two different information sources: (i) the direct interactions among agents and (ii) the information provided by members of the society about experiences they had in the past [35, 36, 41, 42]. Those systems, however, forget a third source of information that can be very useful. As a direct consequence of the interactions, it is possible (even in not too complex societies) to identify different types of social relations between society members. Sociologists and psychologists have been studying these social networks in human societies for a long time and also how these social networks can be used to analyse trust and reputation [33, 5]. These studies show that it is possible to say a lot about the behaviour of individuals using the information obtained from the analysis of their social network.

ReGreT is a modular trust and reputation model oriented to complex e-commerce environments where social relations play an important role. ReGreT model is very relevant in open communities like OpenKnowledge where a social relation exists among the participating agents or one gets gradually built over a period of time. For instance, in the context of a community that collaborate over information, the social relation may be based on roles such as information providers, information brokers, etc. Over time, a role becomes more trust worthy than another, and certain trust/reputation value gets associated with agents holding that role. In some of the human societies, for instance, consider politicians to be less trustworthy than people holding other responsibilities. The ReGreT system attempts to model these emerging trust values associated with being a particular member of the society, and being associated with certain others. Here we brief the main characteristics of ReGreT:

- It takes into account direct experiences, information from third party agents and social structures to calculate trust, reputation and credibility values.
- It has a trust model based on direct experiences and reputation.
- It incorporates an advanced reputation model that works with transmitted and social knowledge.
- It has a credibility module to evaluate the truthfulness of information received from third party agents.
- It uses social network analysis to improve the knowledge about the surrounding society (specially when no direct experiences are available).
- It provides a degree of reliability for the trust, reputation and credibility values that helps the agent to decide if it is sensible or not to use them in the agent's decision making process.
- It can adapt to situations of partial information and improve gradually its accuracy when new information becomes available.

- It can manage at the same time different trust and reputation values associated to different behavioural aspects. Also it can combine reputation and trust values linked to simple aspects in order to calculate values associated to more complex attributes.

The following subsections provide an overview of social network analysis(SNA) and why it can be used to calculate trust and reputation values, when used in complex agent societies. Then a general perspective of the ReGreT system is outlined along with the different elements that compound it.

3.1 Social Network Analysis and agent societies

Social network analysis is the mapping and measuring of relationships between people, groups, organizations, computers or other information/knowledge processing entities. The nodes in the network are the people and groups while the links show relationships between nodes. Social network analysis provides both a visual and a mathematical analysis of these relationships.

As pointed out by Scott [37], three main traditions have contributed to the development of present-day social network analysis: the advances on graph theory performed by the sociometric analysts; the Harvard researchers of the 1930s, who explored patterns of interpersonal relations and the formation of 'cliques'; and the Manchester anthropologists, who built on both of these strands to investigate the structure of 'community' relations in tribal and village societies. These traditions were brought together in the 1960s and 1970s to forge contemporary social network analysis. From then, social network analysis has been widely used in the social and behavioral sciences, as well as areas like political science, economics, or industrial engineering.

One of the main characteristics of social network analysis is the use of relational data instead of attribute data (which is usually quantified and analysed through statistical methods). Relational data can be handled and managed in matrix form or using graphs. A graph structure that shows social relations is called a *sociogram*. A different sociogram is usually built for each social relation under study and depending on the type of relation we have a directed or non-directed sociogram, with weighted edges or without. Indegree, density or node centrality are examples of graph theory concepts used in social network analysis to extract conclusions from sociograms.

The ReGreT system uses social network analysis in two different situations. One is to choose a good set of witnesses to be queried for information. The way social network analysis is used here has a lot of aspects in common with the way it is used in the work of Buskens and Pujol et al. In this situation it is considered only one type of relation and the analysis is based on parameters like centrality, the number of other points in its neighbourhood (degree) and so on. However, in the ReGreT system, social network analysis is also used as part of the reputation and credibility models. In both cases only the relations among a small set of individuals are considered and the type of relation is very relevant in order to perform the analysis.

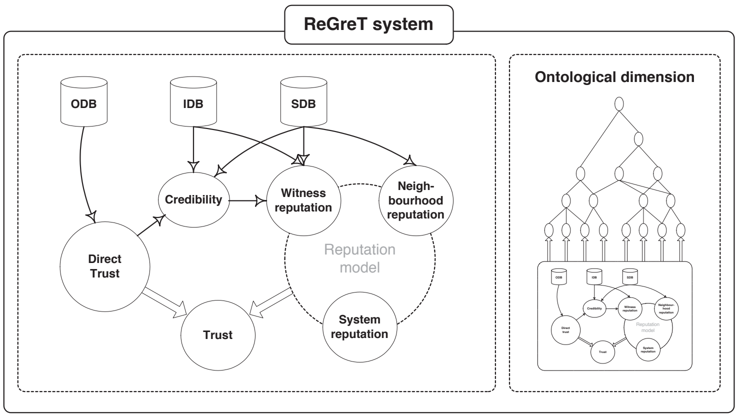


Figure 1: The ReGreT system.

3.2 The ReGreT system, a general view

Figure 1 shows a panoramic view of the ReGreT system.

The system maintains three knowledge bases. The outcomes data base (*ODB*) to store previous contracts and their result; the information data base (*IDB*), that is used as a container for the information received from other partners and finally the sociograms data base (*SDB*) to store the sociograms that define the agent social view of the world. These data bases feed the different modules of the system. They are the *direct trust* module along with the *reputation model*, and the *credibility module*. The reputation model consists of *witness reputation*, *neighbourhood reputation* , and the *system reputation*, each of which will be discussed in the following subsections. The last element which is the *ontological structure* provides the necessary information to combine reputation and trust values linked to simple aspects in order to calculate values associated to more complex attributes.

Trust and reputation have a temporal dimension. That is, the reputation and trust value of an agent change along time. We will, however, omit the reference to time in the notation in order to make it more readable. We will refer to the agent that is calculating a reputation as a (what we call the "source agent") and the agent that is the object of this calculation as b (what we call the "target agent").

3.3 Direct trust

We use the term *direct trust* to refer to the trust that is built from direct interactions. For simplicity, we don't differenciate between direct observation and direct interaction. In the ReGreT system, *direct trust* is always linked to a specific behavioural aspect. Therefore, we talk about the *direct trust* agent a

has in agent b in a specific context to perform a specific action. I can trust a friend to drive me to the airport but it doesn't mean I trust him to fly the plane. The ReGreT system, either for trust or reputation, always takes into account the context.

The basic element to calculate a *direct trust* in the ReGreT system is the *outcome*. We define the *outcome* of a dialog between two agents as either:

- An initial contract to take a particular course of action and the actual result of the actions taken, or

- An initial contract to fix the terms and conditions of a transaction and the actual values of the terms of the transaction.

An outcome is represented as a tuple of the form $o = (a, b, I, X^c, X^f, t)$ where a and b are the agents involved in the contract, I a set of indexes that identify the issues of the contract, X^c and X^f are two vectors with the agreed values of the contract and the actual values after its fulfillment respectively, and t the time when the contract was signed. We use a subscript $i \in I$ to refer to the specific value of issue i in vectors X^c and X^f. For instance, in a SuppWorld scenario we have $I = \{Price, Quantity, Quality, Transport\_Type\}$. If we want to make reference to the *Price* value in the vector X^c we use the notation X^c_{Price}.

ODB is defined as the set of all possible outcomes. $ODB^{a,b} \subseteq ODB$ is the set of outcomes that agent a has signed with agent b. We define $ODB^{a,b}_{\{i_1,\cdots,i_n\}} \subseteq ODB^{a,b}$ as the set of outcomes that include $\{i_1, \cdots, i_n\}$ as issues in the contract. For example, $ODB^{a,b}_{\{Price\}}$ is the set of outcomes that has agent a from previous interactions with agent b and that fix, at least, the value for the issue *Price*.

Given that, we can define a *direct trust* (noted as $DT_{a \to b}(\varphi)$ where φ is the behavioural aspect under evaluation) as the trust relationship calculated directly from an agent's outcomes database.

To calculate a *direct trust* relationship we use a weighted mean of the outcomes evaluation, giving more relevance to recent outcomes.[1] The evaluation of an outcome $o = (a, b, I, X^c, X^f, t)$ (what we call the *impression* of the outcome) depends on the behavioural aspect. This dependency is reflected in two aspects. First, the issue of the outcome that is relevant for the evaluation and second the function used for the evaluation.

We define a *grounding relation* (gr) as the relation that links a behavioural aspect φ with a specific issue and the function used to evaluate the outcome. This allows us to select the right subset of outcomes from the general outcomes' data base and also evaluate the outcome according to the semantics of the behavioural aspect.

As an example, a possible *grounding relation* for a seller in a SuppWorld scenario is defined in the following table:

[1] There are many psychological studies that support recency as a determinant factor [22].

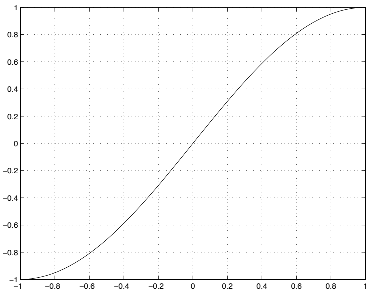


Figure 2: $g(x) = \sin(\frac{\pi}{2}x)$

φ	$gr(\varphi)$	$V(X^s) \otimes V(X^c)$
offers_good_prices	*Price*	$V(X^s) - V(X^c)$
maintains_agreed_quantities	*Quantity*	$abs(V(X^s) - V(X^c))$
offers_good_quality	*Quality*	$V(X^s) - V(X^c)$
delivers_quickly	*Transport_Type*	$V(X^s) - V(X^c)$

where $V(X^c)$ is the utility of the contract values, and $V(X^s)$ is the utility of a vector build using the following formula:

$$X_i^s = \begin{cases} X_i^f & \text{if } i \in gr(\varphi) \\ X_i^c & otherwise \end{cases}$$

In other words, we obtain this vector from vector X^c by replacing the value specified in the index $gr(\varphi)$ by the value in the same positions in vector X^f.

The general formula to evaluate an outcome is:

$$Imp(o, \varphi) = g(V(X^s) \otimes V(X^c))$$

Where g is a function that models the personality of the agent as the degree of deception or reward obtained after the analysis of the outcome (an appropriate function is $g(x) = \sin(\frac{\pi}{2}x)$ shown in figure 2) and $\otimes$ is an aggregation function that depends on the shape of the utility function for that issue. For instance, if instead of having the behavioural aspect *offers_good_prices* we had *offers_bad_prices*, the function $\otimes$ would be $V(X^c) - V(X^s)$.

Given that, the formula to calculate a *direct trust* value in the ReGreT system is:

$$DT_{a \to b}(\varphi) = \sum_{o_i \in ODB^{a,b}_{gr(\varphi)}} \rho(t, t_i) \cdot Imp(o_i, \varphi)$$

with $\rho(t, t_i) = \frac{f(t_i,t)}{\sum_{o_j \in ODB^{a,b}_{gr(\varphi)}} f(t_j,t)}$ where t is the current time and $f(t_i, t)$ is a time dependent function that gives higher values to values closer to t. A simple example of this type of function is $f(t_i, t) = \frac{t_i}{t}$.

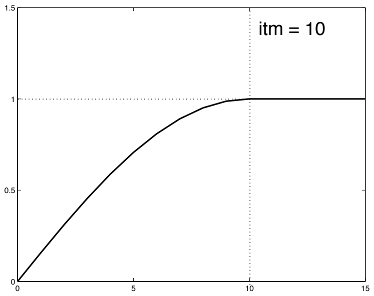


Figure 3: $No(ODB^{a,b}_{gr(\varphi)})$, $itm = 10$

We know how to calculate a *direct trust* value. However in order to use that value it is very important for the agent to know also how reliable it is. There are many elements that can be taken into account to calculate the reliability of a *direct trust* value. The ReGreT system focus on two of them: the number of outcomes used to calculate the *direct trust* value and the variability of their values. This approach is similar to that used in the Sporas reputation model [42].

The intuition behind the number of outcomes factor (noted as No) is that an isolated experience (or a few of them) is not enough to make a correct judgment about somebody. You need a certain amount of experiences before you can assess how an agent behaviour is. As the number of outcomes grows, the reliability degree increases until it reaches a maximum value, what we call the *intimate* level of interactions (*itm* from now on). From a social point of view, this stage is what we know as a close relation. More experiences will not increase the reliability of our opinion from then on. The next simple function is the one we use to model this:

$$No(ODB^{a,b}_{gr(\varphi)}) = \begin{cases} \sin\left(\frac{\pi \cdot |ODB^{a,b}_{gr(\varphi)}|}{2 \cdot itm}\right) & |ODB^{a,b}_{gr(\varphi)}| \leq \text{itm} \\ \\ 1 & \text{otherwise} \end{cases}$$

The function chosen to compute $ODB^{a,b}_{gr(\varphi)}$ when $|ODB^{a,b}_{gr(\varphi)}| \leq$ itm serves the purpose of reaching the value 1 when $|ODB^{a,b}_{gr(\varphi)}| =$ itm and 0 when $|ODB^{a,b}_{gr(\varphi)}| =$ 0. Other functions sharing this property could be used as well.

There is nothing special with the equation we use when $|ODB^{a,b}_{gr(\varphi)}| \leq$ itm. The important thing is that arrives to 1 when $x =$ itm. Other equations can be used to model a more credulous or distrustful behaviour.

The *itm* value is domain dependent: it depends on the interaction frequency of the individuals in that society and also on the “quality” of those interactions. A plot of this function when $itm = 10$ is shown in figure 3.

The outcome deviation (noted as Dv) is the other factor that the ReGreT system takes into account to determine the reliability of a *direct trust* relationship. The greater the variability in the rating values the more volatile will the other agent be in the fulfillment of its agreements. To have a measure of this variability we consider the *impressions* of the outcomes that are used to calculate the *direct trust*.

We calculate the outcome reputation deviation as:

$$Dv(ODB^{a,b}_{gr(\varphi)}) = \sum_{o_i} \rho(t, t_i) \cdot |Imp(o_i, \varphi) - DT_{a\to b}(\varphi)|$$

Where $o_i \in ODB^{a,b}_{gr(\varphi)}$ and $Dv(ODB^{a,b}_{gr(\varphi)}) \in [0,1]$. A deviation value near 1 indicates a high variability in the rating values (that is, a low credibility on the *direct trust* value from the outcome reputation deviation point of view) while a value close to 0 indicates a low variability (that is, a high credibility on the *direct trust* value). Note that we are calculating a kind of weighted mean deviation instead of a standard deviation.

Finally, we define the reliability of a *direct trust* relationship value (*DTRL*) as the product of functions *No* and (1-*Dv*).

$$DTRL_{a\to b}(\varphi) = No(ODB^{a,b}_{gr(\varphi)}) \cdot (1 - Dv(ODB^{a,b}_{gr(\varphi)}))$$

3.4 The reputation model

The problem with direct experiences is that they are usually expensive to obtain in terms of time and cost. This aggregation of others' experience is the base of reputation. The reputation model of the ReGreT system differentiates three types of reputation depending on the information source that is used to calculate them: *Witness Reputation*, *Neighbourhood Reputation* and *System Reputation*.

Sociologically speaking, this division is far from complete, nevertheless is enough to keep the balance between the complexity of the system and the demands that an agent can satisfy in an open community setting. In the following subsections we explain in detail how each reputation type is calculated and how the ReGreT reputation model aggregates the information to obtain a single reputation value.

3.4.1 Witness reputation

Beliefs about trust can be (and usually are) shared among members of a society. The reputation that an agent builds on another agent based on the beliefs gathered from society members (witnesses) is what we call *witness reputation*. In an ideal world, with only homogeneous and trusty agents, this information would be as relevant as direct experiences. However, in the kind of scenarios we are considering, it may happen that *Information be wrong*, *Information be biased*, *Agents hide information* due to various factors associated with the agents in question and the context.

Besides that, the information that comes from other agents can be correlated (what is called the *correlated evidence problem* [32]). This happens when the opinions of different witnesses are based on the same event(s) or when there is a considerable amount of shared information that tends to unify the witnesses' way of "thinking". In both cases, the trust on the information shouldn't be as high as the number of similar opinions may suggest. We take an approach based on the social relations between agents. Analysing these relations, an agent can obtain useful information to minimize the effects of the correlated evidence problem.

We assume that the information to be exchanged among agents is a tuple where the first element is the trust value on the target agent for a specific behavioural aspect from the point of view of the witness, and the second element is a value that reflects how confident the witness is about that trust value. We note the tuple as $\langle \mathit{Trust}_{w\to b}(\varphi), \mathit{TrustRL}_{w\to b}(\varphi)\rangle$, where w is the agent giving the information (the witness), b the target agent and φ the behavioural aspect considered. Each agent maintains a data base of received information. Similar to the outcomes data base, IDB^a is defined as the set of all information received by agent a and $\mathit{IDB}^{a,w}$ notes the subset of information received by agent a from agent w.

- **Identifying the witnesses**

  The first step to calculate a witness reputation is to identify the set of witnesses ($\mathbf{W}$) that will be taken into account by the agent to perform the calculation. The initial set of potential witnesses might be the set of all agents that have interacted with the target agent in the past. For instance, in an e-commerce environment, the initial set can be composed by all the agents that had had a trade relation with the target agent (it seems logical to think that the best witnesses about the commercial behaviour of the target agent are those agents that had a trade relation with it before). This set, however, can be very big and the information provided by its members probably suffer from the correlated evidence problem.

  We take the stance that grouping agents with frequent interactions among them and considering each one of these groups as a single source of information minimizes the correlated evidence problem. Moreover, assuming that asking for information has a cost, it makes no sense to ask for the same thing to agents that we expect will give us more or less the same answer. Grouping agents and asking for information to the most representative agent within each group reduces the number of queries to be done. A domain dependent sociogram is what we use to build these groups and to decide who is their most representative agent.

  There are many heuristics that can be used to find groups and to select the best individual to ask. The heuristic used by the ReGreT witness reputation mechanism is based on the work by Hage and Harary [17]. Taking as the initial graph the subset of the selected sociogram over the agents that had interactions with the target agent, and selecting a node which is a cut point(indicating some kind of local centrality).If there are no cut points, then choose a node as representative with the largest degree.

- **Who can I trust? The credibility model**

  Once the information is gathered from witnesses (or recovered from the data base of previous informations -*IDB*-), the agent obtains

  $$\{\langle \mathit{Trust}_{w_i \to b}(\varphi),\, \mathit{TrustRL}_{w_i \to b}(\varphi)\rangle \mid w_i \in \mathbf{W}\}$$

  where $\mathbf{W}$ is the subset of witnesses whom the agent has selected to be its sources of information. The next step is to aggregate these values to obtain a single value for the *witness Reputation*. As we said before, however, it is possible that this information be wrong or biased. The agent has to be careful to give the right degree of reliability to each piece of information. The importance of each piece of information in the final reputation value will be proportional to the witness credibility.

  Two different methods are used to evaluate the witness credibility.

  The first method is based on the social structure among the witness, the target agent and the source agent. The idea is similar to that used to calculate the neighbourhood reputation. We define $\mathit{socialCr}(a, w_i, b)$ as the credibility that agent a gives to w_i when w_i is giving information about b, taking only into account the social relations among a, w_i and b.

  ReGreT uses fuzzy rules [43] to calculate how the structure of social relations influences the credibility on the information. The antecedent of each rule is the type and degree of a social relation (the edges in a sociogram) and the consequent is the credibility of the witness from the point of view of that social relation. For example:

  IF $\mathit{coop}(w_i, b)$ is high
  THEN $\mathit{socialCr}(a, w_i, b)$ is very_low

  that is, if the level of cooperation between w_i and b is high then the credibility that the information coming from w_i related to b has, from the point of view of a, is very low. The heuristic behind this rule is that a cooperative relation implies some degree of complicity between the agents that share this relation so the information coming from one about the other is probably biased.

  Which relations are relevant to calculate the credibility depends on the meaning that each relation type has in the specific agent community. In a *SuppWorld* scenario, for instance, a trade relation cannot cast any light on the credibility of the information coming from the agents involved in that relation (always from the point of view of social analysis). In other scenarios, however, this could be the other way around.

  Following with the SuppWorld scenario, from the set of social relations only the cooperative relation (*coop*) and the competitive (*comp*) relation are relevant to calculate a measure of credibility. Hence, together with the "no relation" (*no_rel*) possibility there are 9 social structures to be considered as shown in Figure 4.

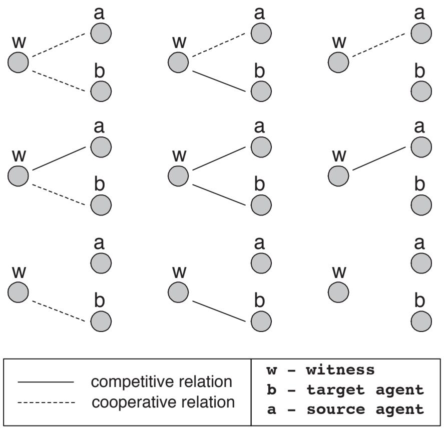


Figure 4: Relevant social structures in a SuppWorld scenario to evaluate credibility.

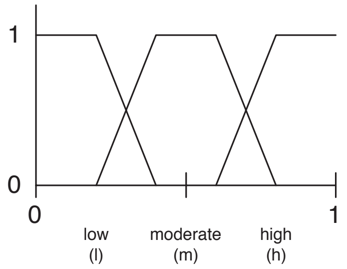


Figure 5: Intensity of a social relation.

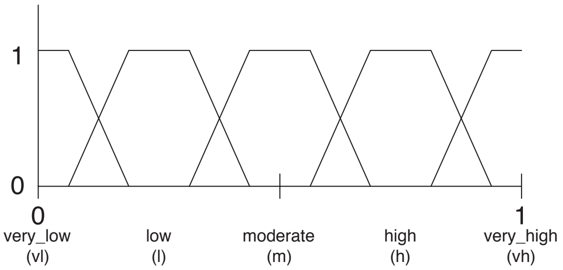


Figure 6: Fuzzy sets for the variable $socialCr(a, w_i, b)$.

Figure 5 shows the fuzzy sets —that give the meaning of the intensity labels used on the arcs of the sociogram— for the values $coop(w_i, a)$, $coop(w_i, b)$, $comp(w_i, a)$, and $comp(w_i, b)$, and figure 6 shows the fuzzy sets for the variable $socialCr(a, w_i, b)$. The variable *no_rel* is boolean. Table 1 shows a possible set of fuzzy rules. Note that a great percentage of the rules tend to be "pessimistic". This is because in those cases where it is not clear that the behaviour is going to be good, we think it is preferable to be cautious. At this moment the kind of influence of each social structure is hand-coded and based on human common sense. An improvement would be the use of a rule learning mechanism to automate the process.

IF	$coop(w_i, a)$ is l	THEN	$socialCr(a, w_i, b)$ is h
IF	$coop(w_i, a)$ is m	THEN	$socialCr(a, w_i, b)$ is vh
IF	$coop(w_i, a)$ is h	THEN	$socialCr(a, w_i, b)$ is vh
IF	$comp(w_i, a)$ is l	THEN	$socialCr(a, w_i, b)$ is m
IF	$comp(w_i, a)$ is m	THEN	$socialCr(a, w_i, b)$ is l
IF	$comp(w_i, a)$ is h	THEN	$socialCr(a, w_i, b)$ is vl
IF	$coop(w_i, b)$ is l	THEN	$socialCr(a, w_i, b)$ is m
IF	$coop(w_i, b)$ is m	THEN	$socialCr(a, w_i, b)$ is l
IF	$coop(w_i, b)$ is h	THEN	$socialCr(a, w_i, b)$ is vl
IF	$comp(w_i, b)$ is l	THEN	$socialCr(a, w_i, b)$ is m
IF	$comp(w_i, b)$ is m	THEN	$socialCr(a, w_i, b)$ is l
IF	$comp(w_i, b)$ is h	THEN	$socialCr(a, w_i, b)$ is vl
IF	$no\_rel(w_i, b)$	AND THEN	$no\_rel(w_i, a)$ $socialCr(a, w_i, b)$ is h

Table 1: Social credibility fuzzy rules.

The second method used in the ReGreT system to calculate the credibility of a witness is to evaluate the accuracy of previous pieces of information sent by that witness to the agent. The agent is using the *direct trust* value (see section 3.3) to measure the truthfulness of the information received from witnesses. For example, an agent a receives information from witness w about agent b saying agent b offers good quality products. Later on, after interacting with agent b, agent a realizes that the products that agent b is selling are horrible. This will be reflected in the value of the *direct_trust* associated to the aspect *offers_good_quality* $\langle DT_{a\rightarrow b}(\mathit{offers\_good\_quality}), DTRL_{a\rightarrow b}(\mathit{offers\_good\_quality})\rangle$. If the *direct_trust* value is low (near -1) it means agent b is offering bad products and therefore that agent w was giving wrong information.

Summarizing, what an agent a is using to evaluate the accuracy of a witness w are pairs of tuples of the form:

$$\begin{array}{l}\langle \mathit{Trust}_{w\rightarrow b}(\varphi), \mathit{TrustRL}_{w\rightarrow b}(\varphi)\rangle \\ \langle DT_{a\rightarrow b}(\varphi), DTRL_{a\rightarrow b}(\varphi)\rangle\end{array}$$

with $b \in B$ and $\varphi \in \Phi$, where B is the set of agents in that society and Φ the set of behavioural aspects.

One important property that has to be remarked about these tuples is that they are not static. They change through time either because the agent collects more direct experiences that modify the perspective it has on the target agent (giving or not more credibility to the witness) or because the agent obtains new information from the witness that overwrites the previous one. Only the most recent information referred to a specific target agent and behavioural aspect from a given witness is stored in the information data base (IDB). Giving the witnesses the opportunity to rectify previous information we are allowing them to correct previous mistakes.

By comparing the trust value assigned by the witness with its own perception of the target agent (represented by the *direct trust* value) the agent obtains the degree of truth of that piece of information. However, there is an important aspect we have not considered up to now. When the trust values are very different but the reliability assigned by the witness is very high and the reliability of the *direct trust* is very low, the agent should decrease the credibility of the witness when it is almost sure that the *direct trust* value the agent has calculated is wrong due to lack of knowledge? What happens if it is the other way around? Is it sensible to decrease the trustworthiness of the witness when the witness itself was giving advice about the weakness of the information by means of the reliability value? Clearly, the method used to evaluate the accuracy of a piece of information has to take much into account the reliability values associated to the trust values in order to decide when the accuracy measure is relevant or not.

There are three main situations the model has to consider:

- $DTRL \approx 0$. The agent does not have enough direct knowledge to judge the truthfulness of what the witness is saying.

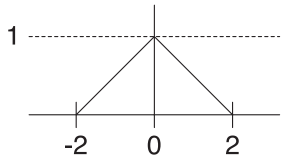


Figure 7: Ap_0 function.

- *TrustRL* ≈ 0. The witness recognizes the weakness of the given information. Therefore that information cannot be used to judge the credibility of the witness.
- *DTRL* ≈ 1, *TrustRL* ≈ 1. The witness is very confident about the information and the agent has enough direct knowledge to judge the truthfulness of that information (and therefore the credibility of the witness).

This can be easily modelled using the product between *TrustRL* and *DTRL* as a factor of relevance for the comparison. Given a piece of information $I = \langle Trust_{w\to b}(\varphi), TrustRL_{w\to b}(\varphi)\rangle \in IDB^{a,w}$, we define the relevance of that information as:

$$\sigma_I = TrustRL_{w\to b}(\varphi) \cdot DTRL_{a\to b}(\varphi)$$

The formula used in the ReGreT system to evaluate the credibility of a witness considering the accuracy of previous information received from that witness is:

$$infoCr(a,w) = \frac{\sum_{I \in IDB^{a,w}_{\sigma>0.5}} \sigma_I \cdot Ap_0(Trust_{w\to b}(\varphi) - DT_{a\to b}(\varphi))}{\sum_{I \in IDB^{a,w}_{\sigma>0.5}} \sigma_I}$$

where $IDB^{a,w}_{\sigma>0.5}$ is defined as $\{I \in IDB^{a,w} : \sigma_I > 0.5\}$. Imposing the restriction of using only those pieces of information with a relevance greater than 0.5 we ensure a minimum quality on the result. The function Ap_0 is depicted in fig 7. If the difference $(Trust_{w\to b}(\varphi) - DT_{a\to b}(\varphi))$ is near 0 it means the witness coincides with the agent (and therefore we have a value for the credibility near 1), on the contrary if the difference shows a value near 2 or -2, it means the witness information is different to what the agent has experienced by itself. The conclusion is that the witness is lying and we obtain value for the credibility of that witness (always associated to that specific piece of information) near 0.

Similar to the case of the *direct trust*, the ReGreT system calculates a measure of reliability for the credibility value *infoCr*. Again, we use the number of values considered for the calculation and the variability of those values as a measure of that reliability. The formula to calculate the reliability of a given *infoCr* value is:

$$\mathit{infoCrRL}(a,w) = \mathit{Ni}(\mathit{IDB}^{a,w}_{\sigma>0.5}) \cdot (1 - \mathit{Dv}(\mathit{IDB}^{a,w}_{\sigma>0.5}))$$

where

$$\mathit{Ni}(\mathit{IDB}^{a,w}_{\sigma>0.5}) = \begin{cases} \sin\left(\frac{\pi \cdot |\mathit{IDB}^{a,w}_{\sigma>0.5}|}{2 \cdot \mathit{itm}}\right) & |\mathit{IDB}^{a,w}_{\sigma>0.5}| \leq \mathrm{itm} \\ 1 & \text{otherwise} \end{cases}$$

$$\mathit{Dv}(\mathit{IDB}^{a,w}_{\sigma>0.5}) = \frac{\sum_{I \in \mathit{IDB}^{a,w}_{\sigma>0.5}} (\sigma_I \cdot |\mathcal{A}|)}{\sum_{I \in \mathit{IDB}^{a,w}_{\sigma>0.5}} \sigma_I}$$

with $\mathcal{A} = \mathit{Ap}_0(\mathit{Trust}_{w\to b}(\varphi) - \mathit{DT}_{a\to b}(\varphi)) - \mathit{infoCr}(a,w)$.

We consider that the credibility calculated considering the accuracy of previous pieces of information (*infoCr*) is more reliable than the credibility based on social relations (*socialCr*). While the analysis of social relations is based on expected behaviours, the analysis of previous information is based on particular facts from the witness the agent wants to evaluate. However, in those situations where there is not enough information to calculate a reliable *infoCr* value, the analysis of social relations can be a good solution. Usually, social relations are easier to obtain than the necessary information to calculate a reliable *infoCr* value. To define the credibility that a witness w_i deserves to an agent a when it is giving information about an agent b we have to differentiate several possibilities:

- Both values (*infoCr* and *socialCr*) can be calculated. The agent uses the *infoCr* value if it is reliable, if not, it uses the credibility based on social relations. The formula for this situation is:

$$\begin{aligned} \mathit{witnessCr}(a,w_i,b) &= \mathit{infoCrRL}(a,w_i) \cdot \mathit{infoCr}(a,w_i) + \\ &\quad (1 - \mathit{infoCrRL}(a,w_i)) \cdot \mathit{socialCr}(a,w_i,b) \end{aligned}$$

- The *socialCr* value is not available (the agent does not have enough social information to calculate it). In this situation we have to differentiate two cases:

$$\begin{array}{rl} \text{If } (\mathit{infoCrRL} > 0.5) & \mathit{witnessCr}(a,w_i,b) = \mathit{infoCr}(a,w_i) \\ \text{Otherwise} & \mathit{witnessCr}(a,w_i,b) = 0.5 \end{array}$$

- The *infoCr* value is not available (the agent does not have direct experiences to evaluate if the information from the witness is reliable or not). Again there are two possibilities:

$$\begin{array}{rl} \text{If } (\mathit{socialCr} \text{ is available}) & \mathit{witnessCr}(a, w_i, b) = \mathit{socialCr}(a, w_i) \\ \text{Otherwise} & \mathit{witnessCr}(a, w_i, b) = 0.5 \end{array}$$

The default value of 0.5 used when there is not enough information to judge the credibility of a witness depends on how credulous the agent is.

- **Witness reputation**

  Now we have all the elements to calculate a *witness reputation* and its associated reliability value considering that the information coming from the witnesses can be wrong or biased. The formulae in the ReGreT system to calculate these values are:

  $$R_{a \overset{W}{\to} b}(\varphi) = \sum_{w_i \in \mathbf{W}} \omega^{w_i b} \cdot \mathit{Trust}_{w_i \to b}(\varphi)$$

  $$RL_{a \overset{W}{\to} b}(\varphi) = \sum_{w_i \in \mathbf{W}} \omega^{w_i b} \cdot \min(\mathit{witnessCr}(a, w_i, b), \mathit{TrustRL}_{w_i \to b}(\varphi))$$

  where $\omega^{w_i b} = \frac{\mathit{witnessCr}(a, w_i, b)}{\sum_{w_j \in \mathbf{W}} \mathit{witnessCr}(a, w_j, b)}$

  These formulae require some explanations. To calculate a *witness reputation* the agent uses the normalized credibility of each witness to weight its opinion in the final value. For the calculation of the reliability, we want that each individual contributes in the same proportion that it has contributed for the calculation of the reputation value. Therefore, the agent uses in the reliability formula the same weights that are used in the reputation formula.

  To calculate the reliability of a witness opinion, the agent uses the minimum between the witness credibility and the reliability value that the witness itself provides. If the witness is a trusty agent, the agent can use the reliability value the witness has proposed. If not, the agent will use the credibility of the witness as a measure for the reliability of the information.

3.4.2 Neighbourhood reputation

The trust on the agents that are in the neighbourhood of the target agent and their relation with it are the elements used to calculate what we call the *Neighbourhood Reputation*. Neighbourhood in a MAS is not related with the physical location of the agents but with the links created through interaction. The main idea is that the behaviour of these neighbours and the kind of relation they have with the target agent can give some clues about the behaviour of the target agent. We note the set of neighbours of agent b as $\mathbf{N}_b = \{n_1, n_2, \cdots, n_n\}$.

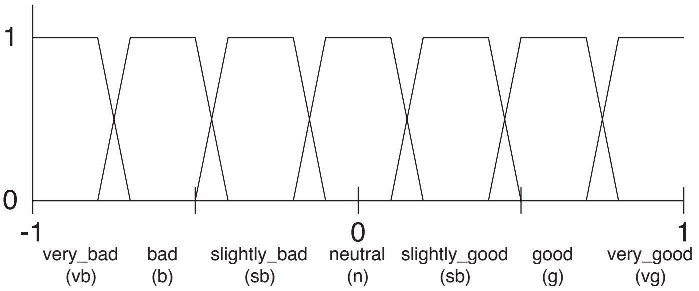


Figure 8: Fuzzy sets for variables $DT_{a \rightarrow n_i}$ and $Rep_{a \xrightarrow{n_i} b}$.

To calculate a *Neighbourhood Reputation* the ReGreT system uses fuzzy rules. The antecedents of these rules are one or several *direct trusts* associated to different behavioural aspects and the relation between the target agent and the neighbour. The consequent is the value for a concrete reputation (that can be associated to the same behavioural aspect of the trust values or not).

The application of these rules generates a set of *individual neighbourhood reputations* noted as $R_{a \xrightarrow{n_i} b}(\varphi)$. For instance, using again the SuppWorld scenario, an example could be:

IF $DT_{a \rightarrow n_i}(\mathit{offers\_good\_quality})$ is X AND $\mathit{coop}(b, n_i) \geqslant$ low
THEN $R_{a \xrightarrow{n_i} b}(\mathit{offers\_good\_quality})$ is X
IF $DTRL_{a \rightarrow n_i}(\mathit{offers\_good\_quality})$ is X' AND $\mathit{coop}(b, n_i)$ is Y'
THEN $RL_{a \xrightarrow{n_i} b}(\mathit{offers\_good\_quality})$ is T(X', Y')

In other words, we are saying that if the neighbour of the target agent is offering good quality products and there is a relation of cooperation between the target and this neighbour, then the target is also assumed to offer good quality products. Here, a neighbour of agent b is an agent that has a *coop* relation with it. The fuzzy sets for variables $DT_{a \rightarrow n_i}$ and $R_{a \xrightarrow{n_i} b}$ are shown in figure 8 and the fuzzy sets for variable $RL_{a \xrightarrow{n_i} b}$ are shown in figure 9.

Finally table 2 shows a possible set of values for function T.

X' Y'	**l**	**m**	**h**
vl	vl	vl	vl
l	vl	vl	l
m	vl	l	m
h	l	m	h
vh	m	h	vh

Table 2: Function T used in reliability rules.

As we have said, instead of relying on the performed actions of the target agent, *Neighbourhood reputation* is using prejudice as a mechanism for evaluation. In human societies the word "prejudice" refers to a negative or hostile

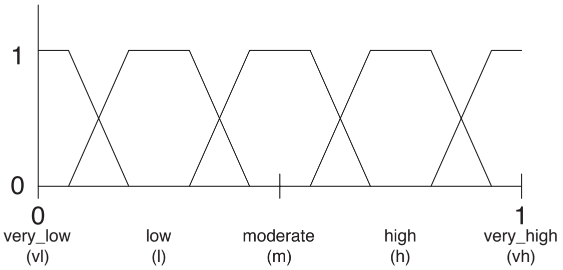


Figure 9: Fuzzy sets for variable $RL_{a \xrightarrow{n_i} b}$

attitude toward another social group, usually racially defined. However we are not talking about human societies where prejudice is without any doubt blameworthy. We are talking about virtual environments populated by software agents. We think that the use of prejudice in the context of agents has a positive aspect. If an agent knows the others are judging it in part because of its partners, it will be careful to choose the right partners and avoid cheaters that perhaps at the beginning can offer better deals but at the end will deteriorate its reputation in front of the community. Moreover, the modular design of the ReGreT reputation model allows to cancel the influence of one type of reputation (in this case the *Neighbourhood reputation*) if it is not useful or convenient in a given environment.

The general formulae we use to calculate a *neighbourhood reputation* and its reliability are similar to those used to calculate a *witness reputation*:

$$R_{a \xrightarrow{N} b}(\varphi) = \sum_{n_i \in N_b} \omega^{n_i b} \cdot R_{a \xrightarrow{n_i} b}(\varphi)$$

$$RL_{a \xrightarrow{N} b}(\varphi) = \sum_{n_i \in N_b} \omega^{n_i b} \cdot RL_{a \xrightarrow{n_i} b}(\varphi)$$

where $\omega^{n_i b} = \frac{RL_{a \xrightarrow{n_i} b}(\varphi)}{\sum_{n_j \in N_b} RL_{a \xrightarrow{n_j} b}(\varphi)}$

In this case we are using the reliability of each *neighbourhood reputation* value to weight the contribution to the final result, both for the reputation and the reliability.

3.4.3 System reputation

The idea behind *System reputations* is to use the common knowledge about *social groups* and the role that the agent is playing in the society as a mechanism to assign default reputations to the agents. We assume that the members of these groups have one or several *observable* features that unambiguously identify their membership. The idea behind *system reputation* is similar to the idea behind *neighbourhood reputation*. As we have seen, *Neighbourhood reputation* focus on reduced groups of agents where the links between their members can not always be easily recognized by the agents that do not belong to the group. On the contrary, the groups considered by *system reputation* are usually mid to big sized and their members can be easily identified. We assume that the role that an agent is playing and the group (or groups) it belongs to is something "visible" and unambiguous for the other agents in that society.

Each time an agent performs an action we consider that it is playing a single role. An agent can play the role of buyer and seller but when it is selling a product only the role of seller is relevant. Although we can think up some situations where an agent can play two or more different roles at a time, we consider that there is always a predominant role and the others can be disregarded.

The knowledge necessary to calculate a *system reputation* is usually inherited from the group or groups to which the agent belongs to. Each group provides knowledge about different aspects. We share the stance that groups influence the point of view of their members [22].

System reputations are calculated using a table for each social group where the rows are the roles the agent can play for that group, and the columns the behavioural aspects.

Table 3 shows an example of *system reputations* for agents that belong to company B from the point of view of an agent of company A. As you notice, in this case the opinion of company A toward agents in company B is not very good.

	offers_good_prices	***offers_good _quality***	***delivers_quickly***	***pays_on_time***
seller	-0.6	-0.8	-0.6	-
buyer	-	-	-	-0.6

Table 3: Example of *system reputations.*

Using a similar table we would define the reliability for these reputations.

System reputations are noted as $R_{a\xrightarrow{S}b}(\varphi)$ and its reliability as $RL_{a\xrightarrow{S}b}(\varphi)$. Hence, for example, using the table defined above, we have that $R_{a\xrightarrow{S}b}(pays\_on\_time) = -0.6$ where b is a buyer that belongs to company B.

The degree of influence that the group or groups to which the agent belongs to have on it, will fix the reliability assigned to *system reputation.*

3.4.4 Combining reputation types

In the previous section we have gone through the three different reputation types considered in the ReGreT reputation model. To these reputation types we have to add a fourth one, the reputation assigned to a third party agent when there is no information at all: the *default* reputation. This reputation is noted as $R_{a\xrightarrow{D}b}(\varphi)$. Usually this will be a fixed value for all b and φ values, however we give the possibility to assign different default reputation values depending on the behavioural aspect (for example, in certain situations the agent could be more trusting). Anyway, what is important is that the default reputation is always available. Similarly to other reputation types, there is also a reliability value associated to the default reputation noted as $RL_{a \xrightarrow{D} b}(\varphi)$.

In this section we will show how these reputations are combined to obtain a single reputation value. As we have seen, each reputation type has different characteristics and there are a lot of heuristics that can be used to aggregate the four reputation values to obtain a single and representative reputation value. The heuristic we propose here is based on the default and calculated reliability assigned to each type.

Assuming we have enough information to calculate all the reputation types, we have the stance that *witness reputation* is the first type that should be considered followed by the *neighbourhood reputation*, *system reputation* and finally the *default reputation*. This ranking, however, has to be subordinated to the calculated reliability for each type.

Given that, we define the reputation that an agent a assigns to an agent b associated to certain behavioural aspect φ as:

$$R_{a \to b}(\varphi) = \sum_{i \in \{W,N,S,D\}} \xi_i \cdot R_{a \xrightarrow{i} b}(\varphi)$$

and similarly for reliability:

$$RL_{a \to b}(\varphi) = \sum_{i \in \{W,N,S,D\}} \xi_i \cdot RL_{a \xrightarrow{i} b}(\varphi)$$

Following the ranking we have established before, the factors $\{\xi_W, \xi_N, \xi_S, \xi_D\}$ we use in the general formula are:

$$\begin{aligned}
\xi_W &= RL_{a \xrightarrow{W} b}(\varphi) \\
\xi_N &= RL_{a \xrightarrow{N} b}(\varphi) \cdot (1 - \xi_W) \\
\xi_S &= RL_{a \xrightarrow{S} b}(\varphi) \cdot (1 - \xi_W - \xi_N) \\
\xi_D &= 1 - \xi_W - \xi_N - \xi_S
\end{aligned}$$

That is, we want the agent to give more relevance to the *witness reputation* in detriment of the others. If the *witness reputation* has a low degree of reliability (for instance because the witnesses are not reliable) then the agent will try to use the *neighbourhood reputation.* If the agent has a poor knowledge of the social relationships and as result of that the reliability of the *neighbourhood reputation* is low, it will try to use the *system reputation.* Finally it will use the *default reputation.*

3.5 Putting all together: the trust model

As showed in figure 1 the ReGreT system considers two elements to calculate the trust on an agent: the reputation of that agent and the *direct trust* (that

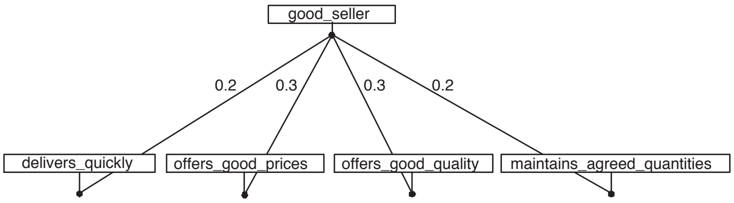


Figure 10: Ontological structure for a buyer in the *SuppWorld* scenario.

is, the result of direct experiences).

As we have argued, *direct trust* is a more reliable source of information than reputation. Using the same approach that for the reputation calculation we define the trust that an agent b deserves to an agent a on certain behavioural aspect φ as:

$$\begin{aligned} \mathit{Trust}_{a \to b}(\varphi) &= \mathit{DTRL}_{a \to b}(\varphi) \cdot \mathit{DT}_{a \to b}(\varphi) + \\ &\quad (1 - \mathit{DTRL}_{a \to b}(\varphi)) \cdot R_{a \to b}(\varphi) \end{aligned}$$

$$\begin{aligned} \mathit{TrustRL}_{a \to b}(\varphi) &= \mathit{DTRL}_{a \to b}(\varphi) \cdot \mathit{DTRL}_{a \to b}(\varphi) + \\ &\quad (1 - \mathit{DTRL}_{a \to b}(\varphi)) \cdot \mathit{RL}_{a \to b}(\varphi) \end{aligned}$$

If the agent has a reliable *direct trust* value, it will use that as a measure of trust. If that value is not so reliable then it will use reputation.

3.6 Ontological dimension

Up to now we have shown how to calculate reputation and trust values linked to behavioural aspects that refer to a single issue of a contract. With the ontological dimension we add the possibility of combining these reputations and trust values associated to simple behaviours to calculate the reputation and trust of more complex behaviours.

To represent the *ontological* dimension we use graph structures. Figure 10 shows an example of a simple ontology structure for a buyer in the *SuppWorld* scenario.

In this case, being a good seller implies delivering products quickly, offering good products, offering good quality and maintain agreed quantities. The buyer gives more relevance to the quality and price of the products to decide if a seller is a good seller or not.

To calculate a given trust taking into account the *ontological dimension*, an agent has to calculate the value of each of the related aspects that, in turn, can be the node of another subgraph with other associated aspects. The trust values for those nodes that are related with an atomic aspect of the behaviour (in the example: *deliver_quickly*, *offer_good_prices*, *offer_good_quality* and *maintain_agreed_quantities*), are calculated using the methods we have presented in the previous sections. Note that an ontology structure can be applied to different parts of the system. The agent can use the ontology either to calculate a reputation value or to calculate a trust value.

The trust over an internal node ψ is computed as follows:

$$Trust_{a \to b}(\psi) = \sum_{\varphi \in children(\psi)} \omega_{\psi\varphi} \cdot Trust_{a \to b}(\varphi)$$

$$TrustRL_{a \to b}(\psi) = \sum_{\varphi \in children(\psi)} \omega_{\psi\varphi} \cdot TrustRL_{a \to b}(\varphi)$$

For instance, using the ontological structure in figure 10 we can calculate the trust on b as a good seller from a's perspective using the formula:

$$\begin{aligned} \mathit{Trust}_{a \to b}(\mathit{good\_seller}) &= 0.3 \cdot \mathit{Trust}_{a \to b}(\mathit{deliver\_quickly}) + \\ &\quad 0.4 \cdot \mathit{Trust}_{a \to b}(\mathit{offer\_good\_prices}) + \\ &\quad 0.4 \cdot \mathit{Trust}_{a \to b}(\mathit{offer\_good\_quality}) + \\ &\quad 0.3 \cdot \mathit{Trust}_{a \to b}(\mathit{maintain\_agreed\_quantities}) \end{aligned}$$

$$\begin{aligned} \mathit{TrustRL}_{a \to b}(\mathit{good\_seller}) &= 0.3 \cdot \mathit{TrustRL}_{a \to b}(\mathit{deliver\_quickly}) + \\ &\quad 0.4 \cdot \mathit{TrustRL}_{a \to b}(\mathit{offer\_good\_prices}) + \\ &\quad 0.4 \cdot \mathit{TrustRL}_{a \to b}(\mathit{offer\_good\_quality}) + \\ &\quad 0.3 \cdot \mathit{TrustRL}_{a \to b}(\mathit{maintain\_agreed\_quantities}) \end{aligned}$$

The same ontological structure could be used to calculate the reputation of being a *good_seller*.

Note that the importance ($\omega_{\psi\varphi}$) of each aspect is agent dependent and not necessarily static. The agent can change these values according to its mental state.

4 CREDIT

CREDIT is a computational trust model (**C**onfidence and **RE**putation **D**efining **I**nteraction-based **T**rust) that is similar to the previous model. It combines *confidence*, on an agent built from direct interactions and *reputation* that is gathered from the experiences of other agents in the community, gossips or by analyzing signals send by the agent. The difference here the method based on fuzzy sets used to compute these measures. [2]

[2] Fuzzy sets are here used to characterise the inherent imprecision in the perception of the performance of an opponent and to provide agents with a high-level means of assessing the extent to an opponent satisfies the issues of a contract. Thus an opponent may be characterised as having a high degree of membership to the fuzzy set 'delivery-on-time' and a low membership to the fuzzy set 'sells-high-quality' to denote that it is expected to deliver on time and sell goods of relatively poor quality.

The use of norms of the environment is a differentiating factor in evaluating the trust of opponents. In so doing, it prevents agents from trusting those opponents that are only performing well because of the prevailing norms. Further CREDIT allows interacting agents, with different norms, to negotiate those issues for which they have different expected values (guided by the norms) and avoid negotiating over those issues for which they have coherent expectations. This, in turn, minimises losses and saves negotiation time. Finally trust can be used to adjust the stance that an agent takes during negotiation so as to minimize the utility loss incurred when it believes its opponent is likely to defect by different degrees from a signed contract. To summarize, CREDIT not only consists of the basic constructs needed to build meaningful measures of trust, it contains the hooks that allow an agents reasoning mechanism to use measure of trust in trust based negotiation (TBN).

In the context of OpenKnowledge, as we use electronic institutions(a regulated environment) to model services, agents enacting the institutions always needs to abide by the norms set by the institution. Where as the heterogeneity of agents and the openness assumption introduces agents from different social settings and belonging to different groups to interact in an institutional context. Thus a norm based evaluation of trust is very relevant here, as it enables the agents to negotiate those issues that is outside of the institutional norms but comes within the social or group norms. Trust has a different interpretation when used in a normative system, for instance an agent delivering a service on time by abiding an institutional norm may not do so if such a norm is not in place. We can model such aspects of trust using the CREDIT model.

The following subsections describe the CREDIT trust model using confidence, reputation, and norms and provides an analysis of the computational complexity involved in the algorithm used in CREDIT, then how CREDIT can be used to influence interactions and empirically evaluates the properties of CREDIT.

4.1 CREDIT model

Let Ag be the society of agents noted as $\alpha, \beta, ... \in Ag$. A particular group of agents is noted as $G \subseteq Ag$ and each agent can only belong to one group. $\mathcal{T}$ denotes a totally ordered set of time points (sufficiently large to account for all agent interactions) noted as $t_0, t_1, \ldots$, such that $t_i > t_j$ if and only if $i > j$.

4.1.1 Contracts

Contracts are agreements about issues and the values these issues should have. Let $X = \{x, y, z, \ldots\}$ be the set of potential issues to include in a contract, and the domain of values taken by an issue x be noted as D_x. Then, a particular contract, O, is an arbitrary set of issue-value assignments noted as $O = \{x_1 = v_1, x_2 = v_2, ..., x_n = v_n\}$ where $x_i \in X$, $v_i \in D_{x_i}$, and $O \in \mathcal{O}$ which denotes the set of potential contracts. Given an agreed contract, two or more agents all have a subset of the contract to enact. Each subset of the contract allocated to an agent is superscripted by the respective agent identifier such that, for example, in a contract O between α and β, $O^\alpha \cup O^\beta = O$. An agent, α, has a utility function for contracts, noted as $U^\alpha : \mathcal{O} \rightarrow [0,1]$, and for each issue $x \in X(O)$ in a contract noted as $U^\alpha_x : D_x \rightarrow [0,1]$. Here the utility of a contract, for an agent is defined, as an aggregation of the weighted utilities of the individual issues as shown below (note this assumes that issues are independent):

$$U^\alpha(O) = \sum_{x \in X(O)} \omega_x \cdot U^\alpha_x(v_x) \tag{1}$$

where $\sum \omega_x = 1$ and $v_x \in D_x$ is the value taken by the issue $x \in X(O)$. Here it is considered that agents, whether from the same group or from different groups, invariably interact within some institutional norms(e.g electronic institutions [13]). CREDIT take into account three general set of norms/rules,(i) *Social rules*, noted as $SocRules$, that all agents in the society Ag possess in common, (ii) *Group rules*, noted as $GroupRules(G)$, that all agents within a particular group $G \subseteq Ag$ have in common, and (iii) *Institutional rules*, noted as $InstRules$, that agents α and β interacting within a particular institution must abide by. Institutional rules may be common to both the interacting agents, where as social rules and group rules are specific to a society or a group and may not be the same for both the interacting agents. Hence its expected utility value should be highlighted. Irrespective of the classification, rules allow an agent to infer expected issue-value assignments from a contract. Here the rules will be written in the following way:

If $x_1 = v_1$ *and* $x_2 = v_2$ *and ... and* $x_m = v_m$ **Then** $x = v$

meaning that if $(x_i = v_i) \in O$ for all $i = 1, \ldots, m$, then issue x's value is expected to be equal to v. The unspecified expectations due to the social setting, O^α_{exp}, of issue-value assignments from O is the set of all conclusions of the rules of agent α, $Rules(\alpha) = SocRules \cup GroupRules(G_1)$ and $InstRules$ (that apply to α and β), that have their premise satisfied by the equalities in the contract O. We can expand the contract O, with the above expectations. The issues contained in the expanded contract may vary (for the same contract O) depending on the group and institutional rules that apply at the time the agents make an agreement. This is because an agent may interact under different institutions (having different institutional norms) or an agent may decide to switch groups to one that has different norms from its original group. Given the expanded contract, an agent may then decide to trust its opponent depending on its prior knowledge of its opponent's performance.

4.2 Confidence

In CREDIT, confidence is defined as

> α's confidence in an issue x handled by β is a measure of certainty (leading to trust), based on evidence from past direct interactions

with β, which allows α to expect a given set of values to be achieved by β for x.

Thus if α has a high degree of confidence with respect to x then it will know what value β is likely to return for x. This value can of course be good for α (give it high utility) or bad (give it low utility). These measures of imprecision on an opponent's behaviour are not strictly probabilistic in nature since they may involve a subjective appreciation of performance as well. Given this, CREDIT takes a fuzzy set based approach to relate confidence levels with expected values for issues.

Agent α's confidence level is defined as the membership level, measured over $[0,1]$, of the behaviour of a particular agent β with respect to an issue x to a linguistic term L, noted as $C(\beta, x, L)$. The cut of the fuzzy set defined by $C(x, L)$ represents a range (on the horizontal axis) of values:

$$E\Delta U_c(x, L) = \{\delta u \in [-1, 1] \mid \mu_L(\delta u) \geq C(x, L)\} \tag{2}$$

that is understood as the range of expected utility deviations at execution time on issue x by agent β. For instance, α may express its belief that β is 'Good' to a confidence level 0.6 in fulfilling the contractual values on price, 'Average' to a level of 0.25, and 'Bad' to a level of 0. This would mean that α expects the utility deviation to lie within the range of values which support the confidence level of 0.6 for 'Good', 0.25 for 'Average', and 0 for 'Bad'.

4.2.1 Evaluating Confidence

Given a a proposed (not yet agreed) contract O, for each issue x in $X(O)$, we can estimate, from the history of past interactions, a probabilistic distribution P of α's utility variation $\Delta U_x \in [-1, 1]$ (negative or positive) relative to issue x. Values of ΔU_x correspond to the possible differences between the utility $U_x(v)$ of the agreed value $(x = v) \in O$ and the utility $U_x(v')$ of the (unknown) final value $(x = v')$ in the executed contract O' (i.e. $\Delta U_x = U_x(v) - U_x(v')$). Then we can say that the agent α has a certain *risk* with issue x when it estimates that $1 \geq q > 0$ where q is the probability that $\Delta U_x < 0$. Of course, the more positive the mean, $\overline{\Delta U}_x$, of this probability distribution (i.e. the higher the expected utility loss), the higher the risk, and the more positive this mean is, the lower the risk (i.e. the lower the expected utility loss).

Now, assume we have a probability distribution P for ΔU_x. In order to determine confidence levels $C(x, L)$ we initially need to determine a significantly representative interval $[\delta_1, \delta_2]$ for ΔU_x (e.g. such that the probability that $(\delta_1 \leq \overline{\Delta U}_x \leq \delta_2)$ is equal to 0.95).

Finally, to calculate confidence levels $C(x, L)$ for each label $L \in \mathcal{L}$, we want the interval $[\delta_1, \delta_2]$ to coincide as much as possible with the set of expected values $E\Delta U_c(x)$ as computed in equation 2. Since this range is defined by the confidence levels of its limits, the procedure amounts to selecting the minimum confidence levels of the two limits for that label as shown in equation 3.

$$C(x, L) = \min(\mu_L(\delta_1), \mu_L(\delta_2)) \tag{3}$$

4.3 Reputation

An agent's reputation is the perception of a group or groups of agents in the society about its abilities and attributes. This model assumes that reputation is simply available from a social network that structures the knowledge that each agent has of its neighbours and keeps track of past interactions as per the ReGreT model discussed above. Here reputation is defined as the following:

> α's estimate of β's reputation in handling an issue x is α's measure of certainty (leading to trust), based on the aggregation of confidence measures (for x) provided to it by other agents which have previously interacted with β, which allows α to expect a given set of values to be achieved by β for x

Hence, we assume that an agent α possesses a function $Rep : Ag \times X \times \mathcal{L} \to [0, 1]$ where $Rep(\beta, x, L)$ represents the reputation of an agent β in handling issue x with respect to the qualifying label L (the name of the agent will be omitted when the context unambiguously determines it). We also assume that the labels $L \in \mathcal{L}$ have their domain specified over the same range of utility deviations (i.e. $\Delta U \in [-1, 1]$).

4.4 Combined Confidence and Reputation Measures

A combination of both measures helps to balance both the societal view on an opponent and the personal view of the agent until the latter can be sure that its own view is more accurate. We assume in this work that the reputation values expressed by each agent in the society represent their confidence values on the behaviour of a given agent. In other words a value $Rep(\beta, x, L)$ represents an aggregation of different confidence values.[3] To come to this conclusion, each agent will have its own threshold on the number of interactions needed to have this accurate measure. Therefore, given agent α's context $\Sigma_{\alpha,\beta} = \langle CB, \{U_x^\alpha\}_{x \in X}, Rules(\alpha), t_c \rangle$, here we propose to define the threshold κ as $\kappa = \max(1, |CB_{\alpha,\beta}|/\theta_{min})$, where $|CB_{\alpha,\beta}|$ is the number of interactions of α with β and θ_{min} is the minimum number of interactions (successful negotiations and completed executions[4]) above which only the direct interaction is taken into account [**?**].

Thus, we capture the combination of confidence and reputation measures through the function $CR : Ag \times X \times \mathcal{L} \to [0, 1]$, which is, in the simplest case,

[3] We are therefore implicitly assuming that all these measures are commensurate (i.e have the same meaning and are based on the same scale), and hence their aggregation make sense.

[4] It is important to specify that only those completed interactions should be taken into account since only these can give us information about the behaviour of the opponent in its execution of contracts. Negotiations could end up in no agreements and these should be excluded when counting interactions in the case base.

a weighted average of both kinds of degrees (as in the previous cases we omit references to the agent whenever possible):

$$CR(x, L) = \kappa \cdot C(x, L) + (1 - \kappa) \cdot Rep(x, L), \tag{4}$$

Given CR levels it is then possible to compute the expected values for an issue x and label L as:

$$E\Delta U_{cr}(x, L) = \{u \mid \mu_L^x(u) \geq CR(x, L)\} \tag{5}$$

and then the intersection of the expected ranges for all the labels $L \in \mathcal{L}$:

$$E\Delta U_{cr}(x) = \bigcap_{L \in \mathcal{L}} E\Delta U_{cr}(x, L) \ . \tag{6}$$

As can be seen, the above range is defined in terms of the utility deviations rather than in terms of the values that the issue could take. However, at negotiation time, for example we might need to compute the expected values an issue could take, after execution of the contract, given an offered value v_0 for the issue. This requires transferring the expected utility deviations to the domain of the issue considered.This can be computed in the following way:

$$EV_{cr}(x, v_0) = \{v \in D_x \mid U_x(v) - U_x(v_0) \in E\Delta U_{cr}(x)\} \tag{7}$$

4.5 Trust

In our trust model we use the combined degrees $\{CR(x, L)\}_{L \in \mathcal{L}}$, as given by equation 4, to define the interval of expected values $E\Delta U_{cr}(x)$, that provides us with a maximum expected loss in utility $\Delta_{loss}^{cr} = \sup(E\Delta U_{cr}(x))$. This maximum expected utility loss represents the risk that is involved in the interaction given knowledge acquired both from direct interactions and reputation and also from the norms of the environment. While the risk describes how much we expect to lose from an interaction, trust is the opposite of this. Thus we define trust as:

$$T(\alpha, \beta, x) = \min(1, 1 - \Delta_{loss}^{cr}) \tag{8}$$

where T serves to describe trust in β for issue x based on both confidence in β and its reputation with respect to issue x.

Here, we choose to bound trust values[5] in the range $[0, 1]$ where 0 represents a completely untrustworthy agent (and corresponds to the maximum possible utility loss) and 1 represents a completely trustworthy agent (and corresponds to zero utility loss).[6]

[5] We acknowledge that other bounds may be applied in other trust models (e.g. $[-1, 1]$ as in [25] or $[0, \infty]$ in eBay). See [25] for a wider discussion on the meaning of the bounds on the rating.

[6] Our choice for the bounds of $[0, 1]$ serves to simplify the analysis when normalising all trust ratings in issues and over contracts.

In any case, we can now define the trust $T(\alpha, \beta, X(O))$ of an agent α in an agent β over a particular set $X(O) = \{x_1, ..., x_k\}$ of issues appearing in the contract O (or in the expanded one O_+) as an aggregation of the trust in each individual issue (e.g. trust in delivering on time, paying on time and the product having the quality specified in the contract). That is, we postulate:

$$T(\alpha, \beta, X(O)) = agg(T(\alpha, \beta, x_1), ..., T(\alpha, \beta, x_k)) \tag{9}$$

where $agg : [0, 1]^k \rightarrow [0, 1]$ is a *suitable* aggregation function[7]. If some issues are considered to be more important than others, the aggregation function should take this into consideration. This can be achieved by means of different weights[8] given for each issue $x_i \in X(O)$ (the higher the weight, the more important the issue). A typical choice would be to take the aggregation[9] function as a weighted mean:

$$T(\alpha, \beta, X(O)) = \sum_{x_i \in X'} w_i \cdot T(\alpha, \beta, x_i) \tag{10}$$

where $\sum w_i = 1$ and $0 \leq w_i \leq 1$.

5 Information based model of trust

This model of trust is based on information theory. It is developed from the observation that any illocutionary exchange between agents give away information, which can be used to build information models of them. Argumentative dialogues change this information model with respect to the ongoing relationship between them. This temporal model builds up trust measures which in turn can be used to select partners for collaboration or to select strategies for argumentation with the chosen partner.

In the context of open communities like OpenKnowledge, where heterogenity of agents is rather the norm, a trust model based on information theory is particularly suited as it assumes almost nothing about the agents, nor about the environment. But builds up a model dynamicaly based on argumentation. It is further based on commitments, thus assumes nothing about the internal architecture of the agents[beliefs, intentions]. Another feature is its honour model, a measure of the integrity of the information exchanged [in appeals] and conditional promises made [in threats and rewards] which supports sustainable partnerships over long periods.

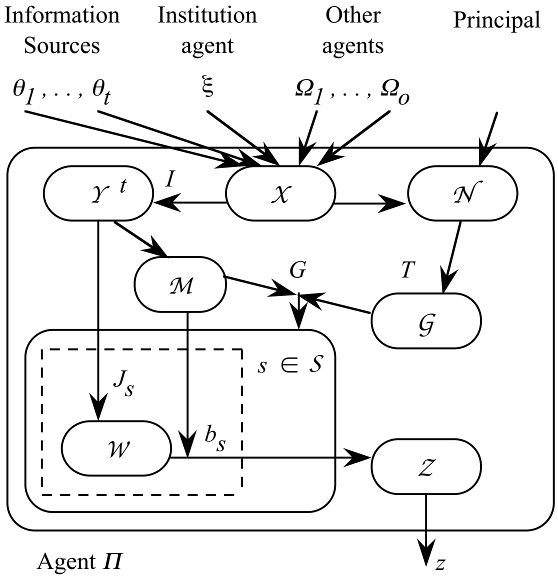
In this section, we discuss the information based agency, information based trust model and a case study discussing the bargaining agent. Discussion on the honour model is not included here, as for the present we are more concerned with individual deals, rather than long term relationships.

[7] Generally, an aggregation function is monotonic such that $\min(u_1, ..., u_k) \leq g(u_1, ..., u_k) \leq \max(u_1, ..., u_k)$ (see [16] for a survey).

[8] Most aggregation operators are defined parametrically with respect to weights assigned to each component to be aggregated (see [16] for more details).

[9] More sophisticated aggregation models (based, for example, on different Lebesgue, Choquet, or Sugeno integrals) could also be used [16].

Figure 11: Basic architecture of agent



The essence of "information-based agency" is described the following. An agent observes events in its environment including what other agents actually do. It chooses to represent some of those observations in its world model as beliefs. As time passes, an agent may not be prepared to accept such beliefs as being "true", and qualifies those representations with epistemic probabilities. Those qualified representations of prior observations are the agent's *information.* This information is primitive — it is the agent's representation of its beliefs about the environment, and about the other agents' prior actions. It is independent of what the agent is trying to achieve, or what the agent believes the other agents are trying to achieve. Given this information, an agent may then choose to adopt goals and strategies, to evaluate situations and to act. If an agent has a utility function that it will have been derived from the agent's information. To enable the agent's strategies to make good use of its information, tools from information theory are applied to summarise and process it. Such an agent is called *information-based.*

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 [4]. Institutions give a normative context to interactions that simplify matters (e.g an agent can't make an offer, have it accepted, and then renege on it). $\delta = (a, b)$ is a deal between two negotiating agents say α and β with α's offer being a and that of β b .

Agent α engages in multi-issue negotiation with a set of other agents: $\{\beta_1, \cdots, \beta_o\}$. The foundation for α's operation is the information that is generated both by and because of its negotiation exchanges. Any message from one agent to another reveals information about the sender. α also acquires information from the environment — including general information sources — to support its actions. α's aim may not be "utility optimisation" — it may not be aware of a utility function. Our approach does not necessarily reject utility optimisation — the selection of a goal and strategy follows the exploitation of the information.

In addition to the information derived from its opponents, α has access to a set of information sources $\{\theta_1, \cdots, \theta_t\}$ that may include the marketplace in which trading takes place, and general information sources such as news-feeds accessed via the Internet. An *institution agent*, ξ, accurately reports to each agents on the execution of commitments, and the fulfilment of promises that involve that agent. The role of the institution agent ξ is simply to "outsource" the agents' observations, so that α is self-contained, complete software agent with the need of an "observe" operation to check whether the fish has arrived from the fishmonger, for example. Together, α, $\{\beta_1, \cdots, \beta_o\}$, ξ, and $\{\theta_1, \cdots, \theta_t\}$ make up a multiagent system.

α has two languages: $\mathcal{C}$ and $\mathcal{L}$ — these are described in Sec. **??**. $\mathcal{C}$ is an illocutionary-based language for communication. $\mathcal{L}$ is a first-order language for internal representation — precisely it is a first-order language with sentence probabilities optionally attached to each sentence representing α's epistemic belief in the truth of that sentence. Fig. 11 shows a high-level view of how α operates. Messages expressed in $\mathcal{C}$ received from $\{\theta_i\}$, ξ and $\{\beta_i\}$ are time-stamped, source-stamped and placed in an *in-box* $\mathcal{X}$. The messages in $\mathcal{X}$ are then translated using an *import function* I into sentences expressed in $\mathcal{L}$ that inherit the time-stamp and source-stamp, they are stored in a *repository* $\mathcal{Y}^t$. The *social model*, $\mathcal{M}$, is a summary of $\mathcal{Y}^t$ and consists of a *trust model* and an *honour model.* And that is all that happens until α triggers a goal.

α triggers a goal, $g \in \mathcal{G}$, in two ways: first in response to a message received from an opponent $\{\beta_i\}$ "I offer you €1 in exchange for an apple", and second in response to some need, $\nu \in \mathcal{N}$, "goodness, we've run out of coffee". In either case, α is motivated by a need — either a need to strike a deal with a particular feature (such as acquiring coffee) or a general need to trade. α's goals could be short-term such as obtaining some information "what is the time?", medium-term such as striking a deal with one of its opponents, or, rather longer-term such as building a (business) relationship with one of its opponents. So α has a trigger mechanism T where: $T : \{\mathcal{X} \cup \mathcal{N}\} \to G$.

For each goal that α commits to, it has a mechanism, G, for selecting a strategy to achieve it where $G : \mathcal{G} \times \mathcal{M} \to \mathcal{S}$ where $\mathcal{S}$ is the strategy library. A *strategy* s maps an information base into an action, $s(\mathcal{Y}^t) = z \in \mathcal{Z}$. Given a goal, g, and the current state of the social model m^t, a strategy: $s = G(g, m^t)$. Each strategy, s, consists of a *plan*, b_s and a *world model* (construction and revision) *function*, J_s, that constructs, and maintains the currency of, the strategy's *world model* W_s^t that consists of a set of probability distributions. A *plan* derives the agent's next action, z, on the basis of the agent's world model for that strategy and the current state of the social model: $z = b_s(W_s^t, m^t)$, and $z = s(\mathcal{Y}^t)$. J_s employs two forms of entropy-based inference:

- Maximum entropy inference, J_s^+, first constructs an *information base* $\mathcal{I}_s^t$ as a set of sentences expressed in $\mathcal{L}$ derived from $\mathcal{Y}^t$, and then from $\mathcal{I}_s^t$ constructs the world model, W_s^t, as a set of complete probability distributions.
- Given a prior world model, W_s^u, where $u < t$, minimum relative entropy inference, J_s^-, first constructs the incremental information base $\mathcal{I}_s^{(u,t)}$ of sentences derived from those in $\mathcal{Y}^t$ that were received between time u and time t, and then from W_s^u and $\mathcal{I}_s^{(u,t)}$ constructs a new world model, W_s^t.

5.1 A commitment based language for Negotiation and Argumentation

In order to express and build negotiation and argumentation dialogues between agents and to internally represent the norms, contracts and commitments, a language needs to be defined. There also needs to be a basic ontology that will be the seed to define this language. This ontology permits to represent the concepts of a given domain, here in this case the domain of trading.

5.1.1 α's Ontology

In order to define the languages that permit the modeling of agent dialogues, we need an ontology that includes a (minimum) repertoire of elements: a set of concepts organized in an *is-a hierarchy*, captured by a partial order relation, and a set of *relations* defined over these concepts. We model ontologies following an algebraic approach [21] as:

An ontology is a tuple $\mathcal{O} = (C, R, \leq, \sigma)$ 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. $\sigma : R \rightarrow C^+$ 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, 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 [23] 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:

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

where l is the length (i.e. number of hops) of the shortest path between the concepts, 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.

5.1.2 α's Languages

Agent α is in a negotiation or a trading relationship with an agent β. They aim to strike a deal $\delta = (a, b)$ where a is α's commitment and b is β's. We denote by A the set of all possible commitments by α, and by B the set of all possible commitments by β. The agents have two languages, $\mathcal{C}$ for communication (illocutionary based) and $\mathcal{L}$ for internal representation (as a restricted first-order language).[10]

The illocutionary particles that support negotiation and argumentation include:

$$\iota = \{\text{Offer}, \text{Accept}, \text{Reject}, \text{Withdraw}, \text{Inform}, \\ \text{Reward}, \text{Threat}, \text{Appeal}\}$$

with the following syntax and informal meaning:
– Offer(α, β, δ) Agent α offers agent β a deal $\delta = (a, b)$ with action commitments a for α and b for β.
– Accept(α, β, δ) Agent α accepts agent β's previously offered deal δ.
– Reject$(\alpha, \beta, \delta, [\mathit{info}])$ Agent α rejects agent β's previously offered deal δ. Optionally, information $[\mathit{info}]$ explaining the reason for the rejection can be given.
– Withdraw$(\alpha, \beta, [\mathit{info}])$ Agent α breaks down negotiation with β. Extra $[\mathit{info}]$ justifying the withdrawal may be given.
– Inform$(\alpha, \beta, \mathit{info})$ Agent α informs β about *info* and commits to the truth of *info*.

[10] It is commonly accepted since the works by Austin and Searle that illocutionary acts are actions that succeed or fail. We will abuse notation in this paper and will consider that they are predicates in a first order logic meaning 'the action has been performed'. For those more pure-minded an alternative is to consider dynamic logic.

– Reward($\alpha, \beta, \delta, \phi, [\mathit{info}]$) Intended to make the opponent accept a proposal with the promise of a future compensation. Agent α offers agent β a deal δ. In case β accepts the proposal, α commits to make ϕ true. The intended meaning is that α believes that worlds in which ϕ is true are somehow desired by β. Optionally, additional information in support of the deal can be given.
– Threat($\alpha, \beta, \delta, \phi, [\mathit{info}]$) Intended to make the opponent accept a proposal with the menace of some sort of retaliation. Agent α offers agent β a deal δ. In case β does not accept the proposal, α commits to make ϕ true. The intended meaning is that α believes that worlds in which ϕ is true are somehow not desired by β. Optionally, additional information in support of the deal can be given.
– Appeal($\alpha, \beta, \delta, \mathit{info}$) Intended to make the opponent accept a proposal as a consequence of the belief update that the accompanying information might bring about. Agent α offers agent β a deal δ. Additionally, α passes a pack of information in support of the deal. An Appeal can be understood as a combination of an offer and an inform, that is Appeal($\alpha, \beta, \delta, \mathit{info}$) = Offer($\alpha, \beta, \delta$); Inform($\alpha, \beta, \mathit{info}$) — we borrow ';' from Dynamic Logic to mean action concatenation.
The accompanying information, [*info*], can be of two basic types: (i) referring to the process (plan) used by an agent to solve a problem, or (ii) data (beliefs) of the agent including preferences. When building relationships, agents will therefore try to influence the opponent by changing their processes (plans) or by providing new data.

5.2 An information based trust model for negotiation

The context of a commitment may be negotiations over individual business deals, or argumentation over business relationships that last over a period of time. For Individual business deals, the concept of trust is a measure of expected deviation of behavior in executing a commitment. Precisely its a measure of how uncertain the enactment of a commitment is. In this sense, the higher the trust the lower the expectation that a significant deviation from what is committed occurs.

Deviations from commitments can occur in two ways. If agent α who is committed to execute a actually executes a' then there are two ways in which a and a' may differ. First, a' may be a variation of commitment a within the ontological context of the negotiation. For example, α may deliver something of slightly inferior quality and, to compensate, deliver an increased quantity of it. Second, the contract variation may involve something outside the ontological context. For example, if α is unable to deliver the quality of wine that was agreed she may "throw in" a box cigars. A contract execution could involve variations of both of these types. In the following we are primarily interested in variations of the first type.

We describe three components that may somehow be combined to describe our trust in an opponent:

- Trust as expected behaviour. The trust I have in an opponent is determined by my evaluation of how he behaves in comparison to my expectations.

- Trust as expected acceptability. The trust I have in an opponent is determined by my evaluation of how good his contract executions are. To capture this we need to define what is meant by saying that a contract execution, b', is better or worse than the signed contract b. We assume here that α's acceptability, $\mathbb{P}^t(\text{IAcc}(\alpha, \beta, \nu, b))$, is fixed from the time of signing the contract to the time of the contract execution. So, if at the time of signing the contract $\mathbb{P}^t(\text{IAcc}(\alpha, \beta, \nu, b')) > \mathbb{P}^t(\text{IAcc}(\alpha, \beta, \nu, b))$ then b' will be preferred to b at the time of contract execution. Similarly for those that are not preferred.

- Trust as certainty in contract execution. The trust I have in an opponent is determined by how consistent he is in the way that he delivers acceptable contract executions.

5.2.1 Information theoretic basis for negotiation

We ground our argumentation model on information-based concepts. *Entropy*, H, is a measure of uncertainty [24] in a probability distribution for a discrete random variable X: $H(X) \triangleq -\sum_i p(x_i) \log p(x_i)$ where $p(x_i) = P(X = x_i)$. Maximum entropy inference and minimum relative entropy inference are chosen partly because of their encapsulation of common sense reasoning [30].

Maximum entropy inference is used to derive sentence probabilities for that which is not known by constructing the "maximally noncommittal" [19] probability distribution, and minimum relative entropy inference is used to update these distributions. These forms of inference are criticised [18] for their dependence on the representation chosen — such as the way in which values for a continuous variable are represented as intervals. We argue to the contrary, that this choice enables the tailoring of the model in fine detail.

Let $\mathcal{G}$ be the set of all positive ground literals that can be constructed using our language $\mathcal{L}$. A *possible world* is a valuation function: $\mathcal{G} \rightarrow \{\top, \bot\}$. $\mathcal{V}|\mathcal{K}^t$ denotes the set of possible worlds that are consistent with an agent's knowledge base $\mathcal{K}^t$ at time t that contains statements which the agent believes are true. A *random world* for $\mathcal{K}^t$ is a probability distribution $W|\mathcal{K}^t = \{p_i\}$ over $\mathcal{V}|\mathcal{K}^t = \{\mathcal{V}_i\}$, where p_i expresses an agent's degree of belief that the possible world $\mathcal{V}_i$ is the actual world. The *derived sentence probability* of any $\sigma \in \mathcal{L}$, *with respect to* a random world $W|\mathcal{K}^t$ is:

$$(\forall \sigma \in \mathcal{L}) P_{W|\mathcal{K}^t}(\sigma) \triangleq \sum_n \{\, p_n \ : \ \sigma \, is \top in \, \mathcal{V}_n \,\}$$

The agent's *belief set* $\mathcal{B}^t = \{\beta_j\}_{j=1}^M$ contains statements to which the agent attaches sentence probabilities $B(\cdot)$. A random world $W|\mathcal{K}^t$ is *consistent* with $\mathcal{B}^t$ if: $(\forall \beta \in \mathcal{B}^t)(B(\beta) = P_{W|\mathcal{K}^t}(\beta))$. Let $\{p_i\} = \overline{W}|\{\mathcal{K}_s^t, \mathcal{B}_s^t\}$ be the "maximum entropy probability distribution over $\mathcal{V}|\mathcal{K}_s^t$ that is consistent with $\mathcal{B}_s^t$". Given an agent with $\mathcal{K}_s^t$ and $\mathcal{B}_s^t$, *maximum entropy inference* states that the *derived sentence probability* for any sentence, $\sigma \in \mathcal{L}$, is:

$$(\forall \sigma \in \mathcal{L}) \mathbb{P}_{\overline{W}|\{\mathcal{K}_s^t, \mathcal{B}_s^t\}}(\sigma) \triangleq \sum_n \{\, p_n \;:\; \sigma \; is \top in \; \omega_n \,\} \tag{11}$$

So each belief imposes a linear constraint on the $\{p_i\}$. The maximum entropy distribution: $\arg\max_{\underline{p}} H(\underline{p})$, subject to these linear constraints, is found by introducing Lagrange multipliers.

Given a prior probability distribution $\underline{q} = (q_i)_{i=1}^n$ and a set of constraints, the *principle of minimum relative entropy* chooses the posterior probability distribution $\underline{p} = (p_i)_{i=1}^n$ that has the least relative entropy with respect to $\underline{q}$, and that satisfies the constraints[11].

[9] describes the estimation of both $P(Acc(\alpha, \beta, \delta))$ and the estimation of $P(Acc(\beta, \alpha, \delta))$ which is α's estimate of β's willingness to accept δ. These estimates are derived by applying maximum entropy inference to the observed behaviour of the agents. In the subsequent subsection we'll see how α updates its sentence probabilities for trust($\cdot$) from observation, decay and experience, preferences and social information following receipt of the illocutionary particles

- **Updating trust from decay and experience**

  An important aspect that we want to model is the fact that beliefs 'evaporate' as time goes by. If we don't keep an ongoing relationship, we somehow forget how *good* the opponent was. If I stop buying from my butcher, I'm not sure anymore that he will sell me the 'best' meat. This decay is what justifies a continuous relationship between individuals. In our model, the conditional probabilities should tend to ignorance. If we have the set of observable contracts as $B = \{b_1, b_2, \ldots, b_n\}$ then complete ignorance of the opponent's expected behaviour means that given the opponent commits to b the conditional probability for each observable contract becomes $\frac{1}{n}$ — i.e. the unconstrained maximum entropy distribution. This natural decay of belief is offset by new observations.

  We define the evolution of the probability distribution that supports the previous definition of decay using an equation inspired by pheromone like models [10]:

$$P^{t+1}(b'|b) = \kappa \cdot \left(\frac{1-\rho}{n} + \rho \cdot \left(P^t(b'|b) + \Delta^t P(b'|b) \right) \right) \tag{12}$$

  where κ is a normalisation constant to ensure that the resulting values for $P^{t+1}(b'|b)$ are a probability distribution. This equation models the

[11] Ie: $\arg\min_{\underline{p}} \sum_{i=1}^n p_i \log \frac{p_i}{q_i}$. The principle of minimum relative entropy is a generalization of the principle of maximum entropy. If the prior distribution $\underline{q}$ is uniform, the relative entropy of $\underline{p}$ with respect to $\underline{q}$ differs from $-H(\underline{p})$ only by a constant. So the principle of maximum entropy is equivalent to the principle of minimum relative entropy with a uniform prior distribution.

passage of time for a conveniently large $\rho \in [0, 1]$ and where the term $\Delta^t P(b'|b)$ represents the increment in an instant of time according to the last experienced event as the following possibilities show.

- **Updating trust from enacted commitments** The choice of a negotiating partner is influenced by the evaluation of the enacted commitments. Eg: Accept(α, β, δ) when uttered by α has the meaning that α becomes socially committed to deliver the deal δ — usually by a given deadline.

  Agent α has the opportunity to evaluate the extent to which agent β sticks to his commitments in Accept$(\cdot)$, Reward$(\cdot)$ and Threat$(\cdot)$ illocutions. We base our measure of trust as the negative entropy of the probability distribution of possible outcomes following such a given commitment — trust measures the relationship between commitment and execution of those commitments. In this way, a natural way to base our modelling of trust is on a conditional probability, P^t, between commitment and evaluation of the enacted commitment as

  $$P^t(\text{evaluate}(\varphi') \mid (\varphi))$$

  Similarity based. The question is how to use the observation of a contract execution c' given a signed contract c in the update of the overall probability distribution over the set of all possible contracts. Here we use the idea that given a particular deviation in a region of the space, *similar* deviations should be expected in other regions. The intuition behind the update is that if my butcher has not given me the quality that I expected when I bought lamb chops, then I might expect similar deviations with respect to chicken. This idea is built upon a function $f(x, y)$ that takes into account the difference between acceptance probabilities and similarity between the perception of the execution x of a contract y, that is a contract for which there was an Accept(β, α, y). Thus, after the observation of c' the increment of probability distribution at time $t + 1$ is:

  $$\Delta^t P(b'|b) = (1 - |f(c', c) - f(b', b)|) \tag{13}$$

  where $f(x, y)$ is

  $$f(x, y) = \\ \begin{cases} 1 & \text{if } P^t(\text{Accept}(x)) > P^t(\text{Accept}(y)) \\ \text{Sim}(x, y) & \text{otherwise.} \end{cases}$$

  and where Sim is an appropriate similarity function (reflexive and symmetric) that determines the indistinguishability between the perceived and the committed contract.

  Entropy based. Suppose that α observes the event $(c'|c)$, the entropy based approach estimates $\Delta^t P(b'|b)$ by applying the principle of minimum relative entropy.[12] Let:

$$\left(P^t_C(b_j|b)\right)_{j=1}^n = \arg\min_{\underline{p}} \sum_{i=1}^n p_i \log \frac{p_i}{P^t(b_i|b)} \tag{14}$$

satisfying the constraint C, and $\underline{p} = (p_j)_{j=1}^n$. Then:

$$\Delta^t P(b'|b) = P^t_C(b'|b) - P^t(b'|b) \tag{15}$$

Constraint C is specified as follows in three cases: first when $c = b$, second when $c' = c \neq b$, and third when $c' \neq c \neq b$.

First, if $c = b$ then C is: $P^t_C(b'|b) = P^t(b'|b) + \nu(1 - P^t(b'|b))$, for $\nu \in [0,1]$ — the value of ν represents the strength of α's belief that the probability that $(b'|b)$ will occur at time $t+1$ should increase if $(b'|b)$ occurs at time t.

Second, if $c' = c \neq b$ then constraint C is:

$$P^t_C(b|b) = P^t(b|b) + g_1(b,c)(1 - P^t(b|b))$$

for: $g_1 \in [0,1]$, where $g_1(b,c)$ represents the strength of α's belief that the probability that $(b|b)$ will occur at time $t+1$ should increase if $(c|c)$ occurs at time t.

Third, if $c' \neq c \neq b$ then suppose that c' is preferred to c by α then $h(c',c) = P^t(\text{Accept}(c')) - P^t(\text{Accept}(c)) > 0$. Let $B(b)^+ = \{x \mid h(x,b) > 0\}$, ie: the set of contract executions that α prefers to b. Given a signed contract b, the prior probability that the contract execution will be preferred by α to b is: $p(b)^+ = \sum_{x \in B(b)^+} P^t(x|b)$. After observing $(c'|c)$ we wish to increase the probability that a preferred execution will occur for contract b to: $p(b \mid (c'|c))^+ = p(b)^+ + g_2(b,c,c')(1 - p(b)^+)$, where $g_2(b,c,c')$ represents the strength of α's belief that the probability that execution of contract b at time $t+1$ will be preferred to b should increase if $(c'|c)$ occurs at time t. Constraint C then is: $\sum_{x \in B(b)^+} P^t_C(x|b) = p(b \mid (c'|c))^+$. Similarly, if c' is *not* preferred to c by α then construct $B(b)^-$.

- **Updating trust from preferences**

  [9] describes the application of maximum entropy inference to enable α to estimate $P^t(\text{Accept}(\beta,\alpha,\delta))$ the probability that β will accept deal δ from α in response to α transmitting the illocution $\text{Offer}(\alpha,\beta,\delta)$. This distribution is derived from previously observed $\text{Offer}(\beta,\alpha,\dots)$ and $\text{Reject}(\beta,\alpha,\dots)$

[12] Given a prior probability distribution $\underline{q} = (q_i)_{i=1}^n$ and a set of constraints, the *principle of minimum relative entropy* chooses the posterior probability distribution $\underline{p} = (p_i)_{i=1}^n$ that has the least relative entropy with respect to $\underline{q}$, $\arg\min_{\underline{p}} \sum_{i=1}^n p_i \log \frac{p_i}{q_i}$, and that satisfies the constraints. The principle of minimum relative entropy is a generalization of the principle of maximum entropy. If the prior distribution $\underline{q}$ is uniform, the relative entropy of $\underline{p}$ with respect to $\underline{q}$ differs from $-H(\underline{p})$ only by a constant. So the principle of maximum entropy is equivalent to the principle of minimum relative entropy with a uniform prior distribution.

illocutions received from β — the former indicating readiness to accept and the latter readiness to reject. α may not accept this historic readiness as being definitive now, if so then $P^t(\text{Accept}(\beta, \alpha, \delta))$ is estimated by attaching time-discounted beliefs (as sentence probabilities) to these observations, and then by calculating the maximum entropy distribution subject to those probabilities as constraints.

Suppose that α now receives preference information from β in the form of an $\text{Inform}(\beta, \alpha, [\mathit{info}])$ illocution, and is prepared to accept this information into its belief set $\mathcal{B}$ as a belief with sentence probability p_{info} — this probability may decay in time. How will this new information influence α's estimate of $P^t(\text{Accept}(\beta, \alpha, \delta))$? Preference information induces a partial ordering on the set of deals. If deal δ_1 is preferred by β to deal δ_2 then: if $\text{Accept}(\beta, \alpha, \delta_2)$ α may conclude to certainty p_{info} that $\text{Accept}(\beta, \alpha, \delta_1)$.

In general, "I prefer deals with property Q_1 to those with property Q_2" becomes the following constraint on the $P^t(\text{Accept}(\beta, \alpha, \delta))$ distribution:

$$p_{\mathit{info}} = \frac{\sum_{\delta:Q_1(\delta)} p_\delta}{\left(\sum_{\delta:Q_1(\delta)} p_\delta\right) + \left(\sum_{\delta:Q_2(\delta)} p_\delta\right)}$$

the posterior distribution for $P^t(\text{Accept}(\beta, \alpha, \delta))$ is calculated by applying the principle of minimum relative entropy[12] to it subject to this constraint.

The method of representing preference information above is quite general. Although if it is used to represent a preference ordering on an issue such as "β prefers to pay less money to more" it generates a set of constraints. If however such a constraint is assumed with $p_{\mathit{info}} = 1$ — ie: if it is represented in the knowledge base $\mathcal{K}$ — then the following device is very economical. [9] describes the representation of $P^t(\text{Accept}(\beta, \alpha, \delta))$ where β is attempting to purchase something for money but with a period of warranty. There α assumes that β prefers "less money to more" and "more warranty to less". These two preference orderings are dealt with neatly by estimating instead $P^t(\text{LimAccept}(\beta, \alpha, \delta))$ meaning "δ is the greatest w.r.t. money and least w.r.t warranty that β will accept from α".

In this way, quantitative preferences over finite domains will give a finite set of linear constraints (in particular, the device above may be used to great effect when $p_{\mathit{info}} = 1$), and qualitative preferences including conditional preferences also yield a finite set of linear constraints.

- **Updating trust from social information**

  Social relationships between agents, and social roles or positions held by agents, introduce a bias, i.e. a constraint, on the admissible probability distributions. A social model can be then a set of constraints introduced in $\mathcal{K}$ that has to be respected by the inference mechanism.

  For instance, with respect to *power*, and assuming we model power as a function from agents to real values, we could model a meek agent by adding the following constraint in $\mathcal{K}$ that establishes different degrees of acceptability for proposals according to the power of the proposer:

$$\begin{aligned}&\text{Power}(\beta) > \text{Power}(\gamma) \rightarrow \\ &\qquad P^t(\text{Accept}(\alpha, \beta, \varphi)) > P^t(\text{Accept}(\alpha, \gamma, \varphi))\end{aligned}$$

A similar case can be made for *reputation*, which refers to the institutional endorsement of observed trustworthiness[13]. Power and reputation are different instruments that help an agent to form an *a priori* assessment of the trustworthiness of an unknown opponent, or to modify the assessment of a known one. If α learns that her good friend γ has a high opinion of β then this may cause α to increase her trust in β and to 'tighten up' the distribution $P^t(b', b)$. Likewise, if α learns that β has a high reputation in a respected institution. So it is natural to represent reputation as Reputation(Φ, β) where Φ is an institution name.

If α receives information, Θ, such as Reputation(Φ, β) then Θ will either be a positive influence on α's estimate of $P^t(b', b)$ [written Θ^+], a negative one [Θ^-], or neutral — ie a positive influence on $P^t(b, b)$ [written Θ^0]. If α receives Θ^+ then her estimate of the probability that the execution of contract b will be preferred to b becomes: $p(b \mid \Theta^+)^+ = p(b)^+ + g_3(b, \Theta^+)(1 - p(b)^+)$, where $p(b)^+$ is the prior probability, $g_3(b, \Theta^+)$ represents the strength of α's belief that the probability that execution of contract b at time $t+1$ will be preferred to b should increase given Θ^+ was received at time t. α revises this estimate using the principle of minimum relative entropy (Eqn. 14) subject to the constraint C: $\sum_{x \in B(b)^+} P^t_C(x|b) = p(b \mid \Theta^+)^+$, where $B(b)^+$ is as in Sec. 5.2.1. Similarly, if α receives Θ^- or Θ^0.

5.2.2 Trust as expected acceptability

The notion of trust was expressed in terms of our expected behaviour in an opponent that was defined for each contract specification b. That notion requires that an ideal distribution, $\mathbb{P}^t_I(b'|b, e)$, has to be specified for each b. The specification of ideal distributions may be avoided by considering "expected acceptability" instead of "expected behaviour". The idea is that we trust β if the acceptability of his contract executions are at or marginally above the acceptability of the contract specification, $\mathbb{P}^t(\text{IAcc}(\alpha, \beta, \nu, (a, b)))$. Defining a function:

$$f(x) = \begin{cases} 0 & \text{if } \ x < \mathbb{P}^t(\text{IAcc}(\alpha, \beta, \nu, (a, b))) \\ 1 & \text{if } \ \mathbb{P}^t(\text{IAcc}(\alpha, \beta, \nu, (a, b))) < x < \mathbb{P}^t(\text{IAcc}(\alpha, \beta, \nu, (a, b))) + \epsilon \\ 0 & \text{otherwise} \end{cases}$$

[13] Electronic Institutions [4] warrant, within specific limits, the *bona fides* of the players therein — it is in their interests to report anecdotal evidence of 'good' behaviour beyond those limits.

(or perhaps a similar function with smoother shape) for some small ϵ, then define:

$$T(\alpha,\beta,b) = \sum_{b'\in B} f(\mathbb{P}^t(\text{IAcc}(\alpha,\beta,\nu,(a,b')))) \cdot \mathbb{P}^t_\beta(b'|b)$$

$$T(\alpha,\beta,\Phi) = \frac{\sum_{\{b\in B|\Phi(b)\}} \mathbb{P}^t(b) \cdot \sum_{b'\in B} f(\mathbb{P}^t(\text{IAcc}(\alpha,\beta,\nu,(a,b')))) \cdot \mathbb{P}^t_\beta(b'|b)}{\sum_{\{b\in B|\Phi(b)\}} \mathbb{P}^t(b)}$$

$$T(\alpha,\beta) = \sum_{b\in B} \mathbb{P}^t(b) \cdot \sum_{b'\in B} f(\mathbb{P}^t(\text{IAcc}(\alpha,\beta,\nu,(a,b')))) \cdot \mathbb{P}^t_\beta(b'|b)$$

5.2.3 Trust as certainty in contract execution

Trust is consistency in expected acceptable contract executions, or “the lack of expected uncertainty in those possible executions that are better than the contract specification”. The idea here is that α will trust β more if variations, b', from expectation, b, are not random. The Trust that an agent α has on agent β with respect to the fulfilment of a contract (a,b) is:

$$T(\alpha,\beta,b) = 1 + \frac{1}{B^*} \cdot \sum_{b'\in B} \mathbb{P}^t_+(b'|b) \log \mathbb{P}^t_+(b'|b)$$

where $\mathbb{P}^t_+(b'|b)$ is the normalisation of $\mathbb{P}^t_\beta(b'|b)$ for those values of b' for which $\mathbb{P}^t(\text{IAcc}(\alpha,\beta,\nu,(a,b'))) > \mathbb{P}^t(\text{IAcc}(\alpha,\beta,\nu,(a,b)))$ and zero otherwise, $B(b)^+$ is the set of contract executions that α prefers to b,

$$B^* = \begin{cases} 1 & \text{if } |B(b)^+| = 1 \\ \log |B(b)^+| & \text{otherwise} \end{cases}$$

and β has agreed to execute b, and α systematically observes b'. Given some b' that α does not prefer to b, the trust value will be 0. Trust will tend to 0 when the dispersion of observations is maximal.

As above we aggregate this measure for those deals of a particular type, that is, those that satisfy $\Phi(\cdot)$:

$$\begin{aligned} T(\alpha,\beta,\Phi) &= 1 + \frac{\sum_{\{b\in B|\Phi(b)\}} \left[\mathbb{P}^t(b) \cdot \sum_{b'\in B} \mathbb{P}^t_+(b'|b) \log \mathbb{P}^t_+(b'|b)\right]}{B^* \cdot \sum_{\{b\in B|\Phi(b)\}} \mathbb{P}^t(b)} \\ &= 1 + \frac{\sum_{\{b\in B|\Phi(b)\}} \sum_{b'\in B} \left[\mathbb{P}^t_+(b',b) \log \mathbb{P}^t_+(b'|b)\right]}{B^* \cdot \sum_{\{b\in B|\Phi(b)\}} \mathbb{P}^t(b)} \end{aligned}$$

where $\mathbb{P}^t_\beta(b',b)$ is the joint probability distribution. And, as a general measure of α’s trust on β we naturally use the normalised negative conditional entropy of executed contracts given signed contracts:

$$\begin{aligned} T(\alpha,\beta) &= 1 + \frac{\sum_{b\in B} \left[\mathbb{P}^t(b) \cdot \sum_{b'\in B} \mathbb{P}^t_+(b'|b) \log \mathbb{P}^t_+(b'|b)\right]}{B^*} \\ &= 1 + \frac{\sum_{b\in B} \sum_{b'\in B} \left[\mathbb{P}^t_+(b',b) \log \mathbb{P}^t_+(b'|b)\right]}{B^*} \end{aligned}$$

5.3 Bargaining– An application of trust

Here a simplistic view of bargaining is assumed, without considering the internals of opponents(avoiding speculation and counter speculation), focusing only on what is known for certain — that is: *what* information is contained in the signals received and *when* did those signals arrive. The communication language that a bargaining agent can use in this scenario is restricted to the illocutions : Offer(·), Accept(·), Reject(·) and Withdraw(·). The actions a bargaining agent performs is as follows:

A simple bargaining agent can only do the following:

- select a bargaining partner;
- make an offer;
- accept an offer;
- withdraw from the negotiation.

Each of these is elaborated in the following.

- **Select a bargaining partner** Selecting a bargaining partner can be done with the help of trust measures that the bargaining agent α has from interactions done so far. α can use some ranking criteria to rank the opponents and select partners for negotiation. This need not be too rigorous as a finer trust measure will be used later to determine whether an offer can be accepted or whether an offer need to be made. Other obvious criteria will be the current need and the knowledge of the opponents as to who among the opponents possess the capability to satisfy the need.

- **Making an offer** Here we discuss the strategies for making an offer with and without a breakdown. This is based on the the principal of equitable information revelation.

  Making offers without breakdown: An agent's strategy s is a function of the information $\mathcal{Y}^t$ that it has at time t. Four simple strategies make offers only on the basis of $\mathbb{P}^t(\text{IAcc}(\alpha, \beta, \nu, \delta))$, α's acceptability threshold γ, and $\mathbb{P}^t(\text{UAcc}(\beta, \alpha, \delta))$. The greedy strategy s^+ chooses:

  $$\arg\max_{\delta}\{\mathbb{P}^t(\text{IAcc}(\alpha, \beta, \nu, \delta)) \mid \mathbb{P}^t(\text{UAcc}(\beta, \alpha, \delta)) \gg 0\},$$

  it is appropriate when α believes Ω is desperate to trade. The *expected-acceptability-to-α-optimizing strategy* s^* chooses:

  $$\arg\max_{\delta}\{\mathbb{P}^t(\text{UAcc}(\beta, \alpha, \delta)) \times \mathbb{P}^t(\text{IAcc}(\alpha, \beta, \nu, \delta)) \mid \mathbb{P}^t(\text{IAcc}(\alpha, \beta, \nu, \delta)) \geq \gamma\}$$

  when α is confident and not desperate to trade. The strategy s^- chooses:

  $$\arg\max_{\delta}\{\mathbb{P}^t(\text{UAcc}(\beta, \alpha, \delta)) \mid \mathbb{P}^t(\text{IAcc}(\alpha, \beta, \nu, \delta)) \geq \gamma\}$$

it optimizes the likelihood of trade — when α is keen to trade without compromising its own standards of acceptability.

Making offers with breakdown: A negotiation may break down because one agent is not prepared to continue for some reason. $p_B = 1 - \mathbb{P}(\text{UWithdraw}(\beta, \alpha, 1)|e)$ is the probability that β will quit in the negotiation in the next round. There are three ways in which α models the risk of breakdown. First, p_B is a constant determined exogenously to the negotiation, in which case at any stage in a continuing negotiation the expected number of rounds until breakdown occurs is $\frac{1}{p_B}$. Second, p_B is a monotonic increasing function of time — this attempts to model an impatient opponent. Third, p_B is a monotonic increasing function of $(1 - \mathbb{P}^t(\text{UAcc}(\beta, \alpha, \delta)))$ — this attempts to model an opponent who will react to unattractive offers.

At any stage in a negotiation α may be prepared to gamble on the expectation that β will remain in the game for the next n rounds. This would occur if there is a constant probability of breakdown $p_B = \frac{1}{n}$. Let $\mathcal{Y}^t$ denote α's the information at time t. s is α's strategy, with plan b_s that determines α's action on the basis of α's world model W_s^t as illustrated in Fig. 11. If α offered to trade with β at $s(\mathcal{Y}_1^t)$ then β may accept this offer, but may have also been prepared to settle for terms more favourable than this to α. If α offered to trade at $s(\mathcal{Y}_1^t \cup \{\text{UAcc}(\beta, \alpha, s(\mathcal{Y}_1^t))\})$ then β will either accept this offer or reject it. In the former case trade occurs at more favourable terms than $s(\mathcal{Y}_1^t)$, and in the latter case a useful belief has been acquired: $\mathbb{B}(\text{UAcc}(\beta, \alpha, s(\mathcal{Y}_1^t))) = 0$, and is added to $\mathcal{Y}_1^t$ before calculating the next offer. This process can be applied twice to generate the offer $s(\mathcal{Y}_1^t \cup \{\text{UAcc}(\beta, \alpha, s(\mathcal{Y}_1^t \cup \{\text{UAcc}(\beta, \alpha, s(\mathcal{Y}_1^t))\}))\})$, or any number of times, optimistically working backwards on the assumption that β will remain in the game for n rounds. The strategy $s^{(n)}$, where $s^{(1)} = s^*$ the expected-acceptability-to-α-optimising strategy $s^{(n)}$ is the strategy of working back from $s^{(1)}$ $(n-1)$ times. At each stage $s^{(n)}$ will benefit also from the information in the intervening counter offers presented by β. The strategy $s^{(n)}$ is reasonable for a risk-taking, expected-acceptability-optimising α.

- **Accepting an offer** To accept an offer $\delta = (a, b)$, an agent that is in no particular hurry to close a deal may be reluctant to accept a proposal unless he believes that his opponent may breakdown the negotiation and withdraw. So we first estimate α's belief in the proposition that β will remain in the negotiation for the next n rounds at least.

  β may withdraw if:

  – he believes that he can get better deal elsewhere

  – he believes that the negotiation is not converging — α can address this by making more attractive proposals.

  – he believes that α is not acting fairly — one factor here is the equitable revelation of information.

To decide whether α should accept an offer, α blends three separate estimations: first, a subjective evaluation (the strength of belief that α has in the proposition that the expected outcome of accepting the proposal will satisfy some of her needs), second, an objective evaluation (the strength of belief that α has in the proposition that the proposal is a "fair deal" in the open market, and third an estimate of whether α will be able to meet her commitment a at contract execution time.

Before making this decision α may also prefer to pause and look for additional information that may reduce this uncertainty — from an information-based agent's point of view, negotiation is firstly an information integrity management problem. To illustrate α's problem at this stage, suppose that she wishes to purchase an air ticket from Barcelona to Sydney. The respected Catalan carrier, Sierra Air, offers the ticket for €1,000, and a new budget Australian carrier, DebJet, for €500. The objective market evaluation, Fair$(\cdot)$, is determined by the amount that Barcelona to Sydney tickets have actually been sold for. The subjective evaluation of expected outcome, Satisfy$(\cdot)$, represents what α's expected evaluation of what each β will actually deliver — this could include delayed flights or the airline ceasing to exist by the departure date.

- **Withdrawing from the negotiation** To decide when to withdraw from a negotiation, α considers the following cases. α may withdraw if:
  - he believes that he can get better deal elsewhere– this might happen in situations where α has not explored enough with the bargaining partners. It may also happen in case where α has a better trust measure for another bargaining partner.
  - he believes that the negotiation is not converging
  - he believes that β is not acting fairly

6 Discussion

We have discussed the three models ReGreT, CREDIT and information based model of trust. The focus of each of these models is different, making a possible combination a powerful one. For instance, the information based model is the most sophisticated model of trust and most suited for open systems. Where as the ReGreT model has a very extensive reputation model, based on social network analysis(SNA). And the CREDIT model focuses on the normative nature of the environment, the institution where the interaction is enacted, the social background of agents and the groups they belong to.

OpenKnowledge requires these three aspects of trust, as it assumes an open system, heterogeneity of agents, a normative environment, and a social network. Thus if we can augment the information based trust model with the reputation model of ReGreT, the normative model of CREDIT then we have good model of trust for the open communities like OpenKnowledge that we are exploring here. The future work involves in combining these models and incorporating this in the setting of OpenKnowlewdge.

References

[1] A. Abdul-Rahman and S. Hailes, *Supporting trust in virtual communities*, Proceedings of the Hawaii's International Conference on Systems Sciences, Maui, Hawaii, 2000.

[2] Gregory D. Abowd and Elizabeth D. Mynatt, *Charting past, present, and future research in ubiquitous computing*, ACM Transactions on Computer-Human Interaction **7** (2000), no. 1, 29–58.

[3] Amazon, *Amazon auctions*, http://auctions.amazon.com, 2002.

[4] Josep Lluis Arcos, Marc Esteva, Pablo Noriega, Juan Antonio Rodríguez, and Carles Sierra, *Environment engineering for multiagent systems*, Journal on Engineering Applications of Artificial Intelligence **18** (2005).

[5] V. Buskens, *Social networks and trust*, Ph.D. thesis, Utrecht University, 1999.

[6] J. Carbo, J.M. Molina, and J. Davila, *Trust management through fuzzy reputation*, Int. Journal in Cooperative Information Systems (2002), in–press.

[7] J. Carter, E. Bitting, and A. Ghorbani, *Reputation formalization for an information-sharing multi-agent sytem*, Computational Intelligence **18** (2002), no. 2, 515—534.

[8] C. Castelfranchi and R. Falcone, *Principles of trust for mas: Cognitive anatomy, social importance, and quantification*, Proceedings of the International Conference on Multi- Agent Systems (ICMAS'98),Paris,France, 1998, pp. 72—79.

[9] J. Debenham, *Bargaining with information*, Proceedings Third International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2004 (N.R. Jennings, C. Sierra, L. Sonenberg, and M. Tambe, eds.), ACM Press, New York, July 2004, pp. 664 – 671.

[10] M. Dorigo and T. Stützle, *Ant colony optimization*, MIT Press, Cambridge, MA, 2004.

[11] eBay, *eBay*, http://www.eBay.com, 2002.

[12] B. Esfandiari and S. Chandrasekharan, *On how agents make friends: Mechanisms for trust acquisition*, Proceedings of the Fourth Workshop on Deception, Fraud and Trust in Agent Societies, Montreal, Canada, 2001, pp. 27—34.

[13] M. Esteva, *Electronic institutions: From specification to development*, Ph.D. thesis, Institut d'Investigació en Intel.ligència Artificial (IIIA), Campus UAB, Catalonia, Spain, 2003.

[14] P. Faratin, C. Sierra, and N. Jennings, *Using similarity criteria to make issue trade-offs in automated negotiation*, Journal of Artificial Intelligence **142** (2003), no. 2, 205–237.

[15] Ian Foster and Carl Kesselman (eds.), *The grid: blueprint for a new computing infrastructure*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[16] M. Grabisch, S. A. Orlovski, and R. R. Yager, *Fuzzy aggregation of numerical preferences*, Fuzzy sets in decision analysis, operations research and statistics (1998), 31–68.

[17] P. Hage and F. Harary, *Structural models in anthropology*, Cambridge University Press, 1983.

[18] J.Y. Halpern, *Reasoning about uncertainty*, MIT Press, 2003.

[19] E.T. Jaynes, *Probability theory — the logic of science*, Cambridge University Press, 2003.

[20] Nick Jennings, Peyman Faratin, Alessandro Lomuscio, Simon Parsons, Carles Sierra, and Mike Wooldridge, *Automated negotiation: Prospects, methods and challenges*, International Journal of Group Decision and Negotiation **10** (2001), no. 2, 199–215.

[21] Yannis Kalfoglou and Marco Schorlemmer, *IF-Map: An ontology-mapping method based on information-flow theory*, Journal on Data Semantics I (Stefano Spaccapietra, Sal March, and Karl Aberer, eds.), Lecture Notes in Computer Science, vol. 2800, Springer-Verlag: Heidelberg, Germany, 2003, pp. 98–127.

[22] Marvin Karlins and Herber I.Abelson, *Persuasion, how opinion and attitudes are changed*, Crosby Lockwood & Son, 1970.

[23] Yuhua Li, Zuhair A. Bandar, and David McLean, *An approach for measuring semantic similarity between words using multiple information sources*, IEEE Transactions on Knowledge and Data Engineering **15** (2003), no. 4, 871 – 882.

[24] D. MacKay, *Information theory, inference and learning algorithms*, Cambridge University Press, 2003.

[25] S. Marsh, *Formalising trust as a computational concept*, Ph.D. thesis, Department of Mathematics and Computer Science, University of Stirling, 1994.

[26] J.V. Neumann and O. Morgenstern, *Theory of games and economic behavior*, Princeton University Press, 1944.

[27] OnSale, *OnSale*, http://www.onsale.com, 2002.

[28] M. J. Osborne and A. Rubinstein, *Bargaining and markets*, Academic Press, 1990.

[29] Julian Padget, Onn Shehory, David Parkes, Norman Sadeh, and William E. Walsh (eds.), *Agent-mediated electronic commerce iv. designing mechanisms and systems*, 4th International Workshop on Agent-Mediated Electronic Commerce, AMEC 2002, held in Bologna, Italy in July 2002 during the AAMAS 2002 conference, 2002.

[30] J. Paris, *Common sense and maximum entropy*, Synthese **117** (1999), no. 1, 75 – 93.

[31] D. Parkes and L. Ungar, *An ascending-price generalized vickrey auction*, Tech. report, Division of Engineering and Applied Sciences, Harvard University, 2002.

[32] J. Pearl, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, Morgan Kaufmann, 1988.

[33] J. M. Pujol, R. Sangesa, and J. Delgado, *Web intelligence*, ch. A Ranking Algorithm Based on Graph Topology to Generate Reputation or Relevance, Springer Verlag, 2003.

[34] H. Raiffa, *Negotiation analysis: The science and art of collaborative decision making*, Harvard U.P., 2002.

[35] J. Sabater and C. Sierra, *Regret: A reputation model for gregarious societies*, Proceedings of the Fourth Workshop on Deception, Fraud and Trust in Agent Societies, Montreal, Canada, 2001, pp. 61—69.

[36] M. Schillo, P. Funk, and M. Rovatsos, *Using trust for detecting deceitful agents in artificial societites*, Applied Artificial Intelligence (2000), no. Special Issue on Trust, Deception and Fraud in Agent Societies.

[37] J. Scott, *Social network analysis*, SAGE Publications, 2000.

[38] S. Sen and N. Sajja, *Robustness of reputation-based trust: Booblean case*, Proceedings of the first international joint conference on autonomous agents and multiagent systems (AAMAS-02), Bologna, Italy, 2002, pp. 288—293.

[39] James Hendler Tim Berners-Lee and Ora Lassila, *The semantic web*, Scientific American (2001).

[40] P. Yolum and M.P. Singh, *Achieving trust via service graphs*, Proceedings of the Autonomous Agents and Multi-Agent Systems Workshop on Deception, Fraud and Trust in Agent Societies, Springer-Verlag, 2003.

[41] Bin Yu and Munindar P. Singh, *A social mechanism of reputation management in electronic communities*, Cooperative Information Agents (CIA), Boston, USA, 2000, pp. 154—165.

[42] Giorgios Zacharia, *Collaborative reputation mechanisms for online communities*, Master's thesis, Massachusetts Institute of Technology, September 1999.

[43] L.A. Zadeh, *Fuzzy logic and approximate reasoning*, Synthese **30** (1975), 407–428.