SUBMITTED : ARTIFICIAL INTELLIGENCE REVIEW
22/10/03

# Review on computational trust and reputation models

Jordi Sabater, Carles Sierra
*IIIA - CSIC, Campus UAB, Bellaterra, Barcelona, Spain*

September 18, 2003

**Abstract.** The scientific research in the area of computational mechanisms for trust and reputation in virtual societies is a recent discipline oriented to increase the reliability and performance of electronic communities. Computer science has moved from the paradigm of isolated machines to the paradigm of networks and distributed computing. Likewise, artificial intelligence is quickly moving from the paradigm of isolated and non-situated intelligence to the paradigm of situated, social and collective intelligence. 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. This review wants to offer a panoramic view on current computational trust and reputation models.

**Keywords:** Trust, Reputation

## 1. Introduction

It is out of discussion the importance of trust and reputation in human societies. Therefore, it is not a surprise that several disciplines, each one from a different perspective, have studied and used both concepts. Psychology (Bromley, 1993; Karlins and I.Abelson, 1970), sociology (Buskens, 1998), philosophy (Plato, 1955; Hume, 1975) and economy (Marimon et al., 2000; Celentani et al., 1966) are a good representation of disciplines that have dedicated efforts to the study of trust and reputation. In this review, however, we will focus our attention on another discipline where the study of trust and reputation has acquired a great relevance in the last few years. We are talking about computer science and specifically about the area of distributed AI. Two elements have contributed to substantially increase the interest on trust and reputation in this area: the multi-agent system paradigm and the spectacular evolution of e-commerce.

The study of trust and reputation has many applications in Information and Communication technologies. Trust and reputation systems have been recognized as key factors for successful electronic commerce adoption. These systems are used by intelligent software agents both as a mechanism to search for trustworthy exchange partners and as an incentive in decision-making about whether or not to honour con-

tracts. Reputation is used in electronic markets as a trust-enforcing, deterrent, and incentive mechanism to avoid cheaters and frauds (eBay, 2002; Amazon, 2002; Dellarocas, 2003). Another important area of application in agent technology is teamwork and cooperation (Montaner et al., 2002).

There are not many works that give a general view of trust and reputation from the point of view of computer science. Dellarocas in his article "The digitalization of Word-Of-Mouth: Promise and Challenges of Online Reputation Mechanisms" (Dellarocas, 2003) presents an overview of online reputation mechanisms that are currently used in commercial web sites. In the area of trust, Grandison et al. in their work "A survey of trust in Internet application" (Grandison and Sloman, 2000) examine the various definitions of trust in the literature and provide a working definition of trust for Internet applications. There are also some proposals to establish a typology for reputation (Mui et al., 2002) and trust (McKnight and Chervany, 2002).

In this article we present a selection of computational trust and reputation models that represent a good sample of the current research. We are not trying to be exhaustive but to provide the reader with a panoramic view that allows to understand which are the approaches and lines of interest in the community. Besides the description of the work that has been done, we also remark those aspects that from our point of view are underexplored and require more attention. Although the study of computational trust and reputation models is quite recent, in the last few years a lot of different proposals have appeared. We think that now is a good moment to take a step back and look at the state of the art to analise the advances and also the drawbacks so the roadmap of future research gets more clear.

The structure of the article is as follows. In section 2 we propose a set of relevant aspects to classify current trust and reputation models. These classification aspects have been selected taking into account the characteristics of the current computational models. In section 3 we go through a representative selection of trust and reputation models describing the main characteristics of each model. In section 4, we classify the models presented in section 3 using the criteria detailed in section 2. Finally, section 5 presents a short discussion.

## 2. Classification dimensions

Trust and reputation can be analyzed from different perspectives and can be used in a wide range of situations. This makes the classification of trust and reputation models a difficult task. In this section

we propose a set of aspects with which we classify the current computational trust and reputation models in a clear landscape. As we have said, we focus our attention on computational models. Therefore, the classification dimensions have been selected considering the special characteristics of these kind of models and the environment where they have to evolve.

### 2.1. Paradigm type

Two different paradigms are currently used in computational trust and reputation models: a cognitive approach and a numerical (or game-theoretic) approach. As pointed out in (Esfandiari and Chandrasekharan, 2001), in models based on a cognitive view, trust is made up of underlying beliefs, and trust is a function of the degree of these beliefs. On the opposite side we have the numerical based models. These models do not rely on beliefs nor intentions to model trust and reputation. Trust and reputation are not the result of a mental process of the agent in a cognitive sense but the result of a more pragmatic game with utility functions, probabilities and the evaluation of past interactions. In the other hand, the cognitive approach tries to reproduce the human reasoning mechanisms behind trust and reputation and the process of building trust and reputation on others is as important as its outcome.

### 2.2. Information sources

It is possible to classify trust and reputation models considering the information sources that they take into account to calculate trust and reputation values. Direct experiences and witness information are the "traditional" information sources used by computational trust and reputation models. In addition to that, a few models have recently started to use information associated to the sociological aspects of agents' behaviour.

The kind of information available to an agent depends on its sensorial capabilities. The use of several information sources increases the reliability of the calculated trust and reputation values but also increases the complexity of the model. Moreover, scenarios that allow agents to obtain diverse information demand smarter (and, therefore, more complex) agents.

#### 2.2.1. *Direct experiences*

This is, without doubt, the most relevant and reliable information source for a trust/reputation model. There are two types of direct experiences that an agent can include as part of its knowledge. The first, and used by all the trust and reputation models analyzed in this

review, is the experience based on the direct interaction with the partner. The second is the experience based on the observed interaction of other members of the community. This second type is not so common and restricted to scenarios that are prepared to allow it. Usually, in those models that consider the observation of other partners activity, a certain level of noise in the so obtained information is assumed.

2.2.2. *Witness information*
Witness information (also called word-of-mouth or indirect information) is the information that comes from other members of the community. That information can be based on their own direct experiences or it can be information that they gathered from others. If direct experience is the most reliable source of information for a trust/reputation model, witness information is usually the most abundant. However, it is far more complex for trust and reputation models to use it. The reason is the uncertainty that surrounds this kind of information. It is not strange that witnesses manipulate or hide pieces of information to their own benefit.

2.2.3. *Sociological information*
This information is only available in scenarios where there is a rich interaction between agents. The base for this knowledge are the social relations between agents and the role that these agents are playing in the society. The social relations established between agents in a multi-agent system are usually a simplified reflection of the more complex relations established between their human counterparts. Currently, only a few trust and reputation models use this knowledge applied to agent communities to calculate or improve the calculation of trust and reputation values. These models use techniques like *social network analysis*. Social network analysis is the study of social relationships between individuals in a society that emerged as a set of methods for the analysis of social structures, methods that specifically allow an investigation of the relational aspects of these structures. The use of these methods, therefore, depends on the availability of relational data (Scott, 2000).

Although currently the number of models that take into account this kind of information is reduced, we guess that the increase of complexity in multi-agent systems will make it more and more important in the near future.

2.2.4. *Prejudice*
The use of prejudice to calculate trust and reputation values is another mechanism not very common but present in current trust and reputation models. As most people today use the word, "prejudice" refers

to a negative or hostile attitude toward another social group, usually racially defined. However, the negative connotations that prejudice has in human societies has to be revised when applied to agent communities. Differently from the signifiers used in human societies that range from skin color to sex, the set of signifiers used in computational trust and reputation models are usually out of ethical discussion.

### 2.3. Visibility types

Trust and reputation of an individual can either be seen as a global property shared by all the observers or as a subjective property 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. In the latter case, we cannot talk about the trust/reputation of an individual $x$, we have to talk about the trust/reputation of an individual $x$ from the point of view of an individual $y$.

The position of taking trust and reputation as a global property is common in online reputation mechanisms (see section 3.2). These systems are intended for scenarios with thousands or even millions of users. As pointed out by Dellarocas (Dellarocas, 2003), the size of these scenarios makes repeated interaction between the same set of players unlikely and, therefore, reduces the incentives for players to cooperate on the basis of hoping to develop a profitable relationship.

Take the example of an electronic auction house like those accessible nowadays through Internet. One day, the user wants to buy a book and the next day s/he wants to buy a computer. The intersection between users selling books and users selling computers is probably empty so the few personal experiences accumulated buying books are not useful in the computers' market. Computer sellers are unknown for the user so s/he has to rely on the information that people who bought computers in the past has left in the form of a reputation value. The robustness of these systems relies on the number of opinions available for a given partner. A great number of opinions minimize the risk of single individual biased perceptions.

In models that consider trust and reputation as a global property, the main problem is the lack of personalization of that value. Something that is bad for me could be acceptable for others and the other way around. Although this approach can be acceptable in simple scenarios where it is possible to assign a common "way of thinking" to all members of the community, it is not useful when agents have to deal with more complex and subjective affairs.

The antithesis of these models are the models that consider trust and reputation as a subjective property. Each agent uses its personal experiences and what the other agents have said to it personally, among other things, to build the trust and reputation of each member of the community. These models are indicated for medium and small size environments where agents meet frequently and therefore it is possible to establish strong links among them.

### 2.4. Model's granularity

Is trust/reputation context dependent? If we trust a doctor when she is recommending a medicine it doesn't mean we have to trust her when she is suggesting a bottle of wine. The reputation as a good sportsman does not help if we are looking for a competent scientist. It seems clear that the answer is yes: trust and reputation are context dependent properties. However, adding to computational trust and reputation models the capability to deal with several contexts has a cost in terms of complexity and adds some side effects that are not always necessary or desirable.

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.

One could argue that it is always possible to transform a single-context model into a multi-context one just having different instances of the single-context model, one for each considered context. However, if there is something in trust and reputation environments that is usually scarce, that is the information used to calculate trust and reputation values. So what really gives to a model the category of being a multi-context model is the capability of making a smart use of each piece of information to calculate different trust or reputation values associated to different activities. Identifying the right context for a piece of information or using the same information in several contexts when it is possible are two examples of the capabilities that define a real multi-context model.

Is this always necessary? Certainly not. Nowadays, there are very few computational trust and reputation models that care about the multi-context nature of trust and reputation and even fewer that propose some kind of solution. This is because current models are focused on specific scenarios with very delimited tasks to be performed by the agents. In other words, it is possible to summarize all the agent activities in a single context without losing too much versatility. However, and similarly to what we have mentioned before about the use of sociological information, as the complexity of tasks to be performed by agents will increase in the near future, we may also expect an increase of the importance devoted to this aspect in trust modelling.

### 2.5. Agent behaviour assumptions

The capacity to deal with agents showing different degrees of cheating behaviour is the aspect considered here to establish a classification. We use three levels to categorize trust and reputation models from this point of view according to what we have observed in the analyzed trust and reputation models:

- Level 0. Cheating behaviour is not considered. The model relies on a large number of agents who offer honest ratings to counteract the potential effect of the ratings provided by malicious agents.
- Level 1. The model assumes that agents can hide or bias the information but they never lie.
- Level 2. The model has specific mechanisms to deal with liars.

### 2.6. Type of exchanged information

The classification dimension here is the type of information expected from witnesses. We can establish two big groups. Those models that assume boolean information and those models that deal with continuous measures. Although it seems a simple difference choosing one approach or the other has a great influence in the design of the model. Usually, models that rely on probabilistic methods work with boolean information while those models based on aggregation mechanisms use continuous measures.

### 2.7. Trust/Reputation reliability measure

Is the model providing a measure of how reliable is the trust/reputation value? Sometimes, 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. Some models incorporate mechanisms that provide this kind of information.

## 3. Computational trust and reputation models

A plethora of computational trust and reputation models have appeared in the last few years, each one with its own characteristics and using different technical solutions. In this section we go through a selection of these models, wide enough to provide a panoramic view of the area.

### 3.1. S. Marsh

The trust model proposed by Marsh (Marsh, 1994) is one of the earliest. The model only takes into account direct interaction. It differentiates three types of trust:

- Basic trust. Models the general trusting *disposition* independently of who is the agent that is in front. It is calculated from all the experiences accumulated by the agent. Good experiences lead to a greater disposition to trust, and vice versa. The author uses the notation $T_x^t$ to represent the trust disposition of agent $x$ at time $t$.

- General trust. This is the trust that one agent has on another without taking into account any specific situation. It simply represents general trust on the other agent. It is noted as $T_x(y)^t$ representing the general trust that agent $x$ has on agent $y$ at time $t$.

- Situational trust. This is the amount of trust that one agent has in another taking into account a specific situation. The *utility* of the situation, its *importance* and the 'General trust' are the elements considered in order to calculate the 'Situational trust'. The basic formula used to calculate this type of trust is:

$$T_x(y,\alpha)^t = U_x(\alpha)^t \times I_x(\alpha)^t \times \widehat{T_x(y)^t}$$

where $x$ is the evaluator, $y$ the target agent and $\alpha$ the situation. $U_x(\alpha)^t$ represents the utility $x$ gains from situation $\alpha$, $I_x(\alpha)^t$ is the importance of the situation $\alpha$ for agent $x$ and $\widehat{T_x(y)^t}$ is the estimate of general trust after taking into account all possible relevant data with respect to $T_x(y,\alpha)$ in the past; i.e., if $t$ is the current time, $x$ will aggregate all situations $T_x(y,\sigma)^T$, with $\theta < T < t$ and $\sigma$ similar

or identical to the present situation $\alpha$. $\theta$ and $t$ define the temporal window that the agent is considering. Only the experiences within that window will be taken into account for the aggregation.

In order to define $\widehat{T_x(y)}$ the author proposes three statistical methods: the mean, the maximum and the minimum. Each method is identified with a different type of agent: the optimistic (that takes the maximum trust value from the range of experiences it has had), the pessimistic (that uses the minimum trust value) and the realistic (that calculates the value as a mean using the formula $\widehat{T_x(y)} = \frac{1}{|A|} \sum_{\alpha \in A} T_x(y, \alpha)$, where $A$ is the set of situations similar to the present situation $\alpha$ available in the temporal window).

These trust values are used to help an agent decide if it is worth it or not to cooperate with another agent. Besides trust, the decision mechanism takes into account the importance of the action to be performed, the risk associated to the situation and the perceived competence of the target agent. To calculate the risk and the perceived competence, different types of trust (basic, general and situational) are used.

Finally, the model also introduces the notion of "reciprocation" as a modifier of the trust values. The idea behind reciprocation is that if an agent $x$ had helped an agent $y$ in the past and $y$ responded that time by defecting, the trust $x$ has on $y$ will be reduced (and the other way around).

### 3.2. Online reputation models

eBay (eBay, 2002), Amazon Auctions (Amazon, 2002) and OnSale Exchange (OnSale, 2002) are good examples of online marketplaces that use reputation mechanisms. eBay (eBay, 2002) is one of the world's largest online marketplace with a community of over 50 million registered users. Most items on eBay are sold trough 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 (Amazon, 2002) and OnSale Exchange (OnSale, 2002) use also a mean (in this case of all ratings) to assign a reputation value.

All these models consider reputation as a global property and use a single value that is not dependent on the context. The information source used to build the reputation value is the information that comes from other agents that previously interacted with the target agent (witness information). They do not provide explicit mechanisms to deal

with users that provide false information. A great number of opinions that "dilute" false or biased information is the only way to increase the reliability of the reputation value.

### 3.3. Sporas and Histos

#### 3.3.1. *Sporas*

Sporas (Zacharia, 1999) is an evolved version of the online reputation models presented in 3.2. In this model, only the most recent rating between two users is considered. Another important characteristic is that users with very high reputation values experience much smaller rating changes after each update than users with a low reputation. Using a similar approach to the Glicko (Glickman, 1999) system —a computational method used to evaluate the player's relative strengths in pairwise games—, Sporas incorporates a measure of the reliability of the users' reputation based on the standard deviation of reputation values.

This model has the same general characteristics as the previously commented online reputation mechanisms 3.2. However, it is more robust to changes in the behaviour of a user and the reliability measure improves the usability of the reputation value.

#### 3.3.2. *Histos*

Histos (Zacharia, 1999) was designed as a response to the lack of personalization that Sporas reputation values have. The model can deal with direct information (although in a very simple way) and witness information. Contrary to Sporas, the reputation value is a subjective property assigned particularly by each individual.

The treatment of direct interaction in this reputation model is limited to the use of the most recent experience with the agent that is being evaluated. The strength of the model relies on its use of witness information.

Pairwise ratings are represented as a directed graph where nodes represent agents and edges carry information on the most recent reputation rating given by one agent to another. The root node represents the agent owner of the graph. This structure is similar to the *TrustNet* used by Schillo et al. (Schillo et al., 2000). The reputation of an agent at level $X$ of the graph (with $X > 0$) is calculated recursively as a weighted mean of the rating values that agents in level X-1 gave to that agent. The weights are the reputations of the agents that rate the target agent. As we have seen, the agents who have been rated directly by the agent owner of the graph have a reputation value equal to the rating value. This is the base case of the recursion. The model also

limits the length and number of paths that are taken into account for the calculation. The reputation value does not depend on the context and no special mechanisms are provided to deal with cheaters.

A drawback of this model is the use of the reputation value assigned to a witness also as a measure of its reliability. If an agent is a good seller, this does not mean that it has to be also a reliable witness.

### 3.4. SCHILLO ET AL.

The trust model proposed by Schillo et al. (Schillo et al., 2000) is intended for scenarios where the result of an interaction between two agents (from the point of view of trust) is a boolean impression: good or bad; there are no degrees of satisfaction. More concretely, to make the experiments they propose a Prisoner's dilemma set of games with a partner selection phase. Each agent receives the results of the game it has played plus the information about the games played by a subset of all players (its neighbours). The result of an interaction in this scenario is an impression on the honesty of the partner (if she did what she claimed in the partner selection phase) and which was the behaviour she had according to the normal prisoner's dilemma actions (cooperation or defection). The model is based on probability theory. The formula to calculate the trust that an agent $Q$ deserves to an agent $A$ (that is, the probability that the agent $A$ be honest in the next interaction) is $T(A,Q) = \frac{e}{n}$ where $n$ is the number of observed situations and $e$ the number of times that the target agent was honest.

Complementing the information that results from direct interaction/observation, an agent can interview other agents that it has met before. Each agent uses a different *TrustNet* data structure. A *TrustNet* is a directed graph where nodes represent witnesses and edges carry information on the observations that the parent node agent told the owner of the net (the root node of the *TrustNet*) about the child node agents.

In this model, testimonial evidence from interviews may be brittle, as witnesses may have different motives and may try to deceive agents about their true observation. Thus, every agent is confronted with noise in the information and also with the possibility that the source of information itself is biasing the data.

The answer of witnesses to a query is the set of observed experiences (and not a summary of them). Given that, the authors assume that it is not worth it for witnesses to give *false* information. A witness will not say that a target agent has played dishonest in game $x$ if this was not the case because the inquirer could have observed the same game and, therefore, notice that the witness is lying. Witnesses do not want

to be uncovered by obviously betraying. Therefore, the model assumes that witnesses never lie but that can hide (positive) information in order to make other agents appear less trustworthy. Assuming that negative information will be always reported by witnesses, the problem is reduced to know to what extend those witnesses have biased the reported data (hiding positive observations).

To do that, betraying (hiding information) is modelled as a stochastic process where an agent decides to inform about a positive fact of another agent with probability $p$ and hide that information with probability $(1-p)$. The application of this process can be seen as a Bernoulli-experiment and the repetition of the experiment as a Bernoulli-chain. Probability theory is then used to estimate the hidden amount of positive information. This process can be applied recursively from the target agent through all its ancestors up to the root node of the *TrustNet*.

With all this process, the agent is building for each piece of information an approximation of what the witnesses would have said if they had been completely honest about their information.

As the information from the witnesses comprises the list of observations it can be collated to eliminate the "correlated evidence" problem (Pearl, 1988). This, however, cannot be done for the hidden information. The proposed solution in this case is based on the assumption that the relation of overlapping of the data in reported and non reported (hidden) information is constant.

No information is given about how to combine direct experiences with information coming from witnesses.

The trust value is a subjective property assigned particularly by each individual and it does not depend on the context.

### 3.5. Abdul-Rahman and Hailes

This trust model (Abdul-Rahman and Hailes, 2000) 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. For instance, if the associated tuple of a partner in a given context is $(0, 0, 4, 3)$ the trust assigned to that partner will be *t* (trustworthy) that corresponds to the third position in the tuple. If there is more than one position in the tuple with the maximum value, the model gives an *uncertainty* trust degree according to a table of pattern situations that cover this cases. There are three possible uncertainty values (and the

corresponding patterns) to cover the situations where there are mostly good and some bad, mostly bad and some good and equal amount of good and bad experiences.

This is the only model analyzed where before combining the information that comes from witnesses, the information is adjusted according to previous information coming from that witness and the consequent outcomes that validate that information. For example, suppose $a$ informs to $x$ that $b$ is $vt$ and $x$'s evaluation of its experience with $b$ is merely $t$. Next time that $a$ gives information to $x$, $x$ will adjust the information accordingly before taking it into account.

The problem of this approach is that it is not possible to differentiate those agents that are lying from those agents that are telling the truth but "think" different. Although there are scenarios where this is not important (like the scenario suggested by the authors where agents recommend goods to other agents) it can be a limitation in some scenarios.

In order to combine information, the model gives more relevance to the information coming from those agents with a more similar point of view. That is, it gives more importance to the information that needs to be adjusted very little or, even better, does not need to be adjusted at all because it comes from agents that have a similar perspective in a given context.

Contrarily to other trust models where witness information is merged with direct information to obtain the trust on the specific 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.

### 3.6. Esfandiary and Chandrasekharan

In the trust model proposed by Esfandiari and Chandrasekharan (Esfandiari and Chandrasekharan, 2001), two one-on-one trust acquisition mechanisms are proposed. The first is based on observation. They propose the use of Bayesian networks and to perform the trust acquisition by Bayesian learning. In the simplest case of a known structure and a fully observable Bayesian network, the learning task is reduced to statistical considerations.

The second trust acquisition mechanism is based on interaction. The approach is the same used in (Lashkari et al., 1994). 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.

A simple way to calculate the interaction-based trust during the exploratory stage is using the formula $T_{inter}(A,B) = \frac{\text{number\_of\_correct\_replies}}{\text{total\_number\_of\_replies}}$.

To deal with witness information, each agent builds a directed labeled graph where nodes represent agents and where an $(a,b)$ edge represents the trust value that $a$ has on $b$. Edges are absent if the trust value is unknown. In such a graph, there is the possibility of having cycles that artificially decrease the trust value and different paths that give contradictory values. To solve this problem, instead of using a single value for trust the model uses a trust interval determined by the minimum and maximum value of all paths without cycles that connect two agents.

The authors claim that the calculation of this trust interval is equivalent to the problem of routing in a communication network and, therefore, known distributed algorithms used to solve that problem can be successfully applied to this situation.

To allow a multi-context notion of trust (see section 2.4) the authors propose the use of colored edges, with a color per task or type of trust. Trust would only propagate through edges of the same color.

Finally, the authors propose a trust acquisition mechanism using institutions, what they call *institutionalized trust*. This is similar to the concept of *system reputation* in the ReGreT (Sabater and Sierra, 2002) model. The idea is to exploit the structure in the environment to determine trust values.

No information is given about how to combine the different trust acquisition mechanisms.

### 3.7. Yu and Singh

In the model proposed by Yu and Singh (Yu and Singh, 2001; Yu and Singh, 2002b; Yu and Singh, 2002a), the information stored by an agent about direct interactions is a set of values that reflect the quality of these interactions (what they call quality of service -*QoS*-). Only the most recent experiences with each concrete partner are considered for the calculations. Each agent defines an upper and lower threshold that define the frontier between what are considered *QoS*s ascribed to trustworthy agents, *QoS*s with no clear classification and *QoS*s ascribed to non trustworthy agents. Then, using the historic information together with Dempster-Shafer theory of evidence, an agent can calculate the probability that its partner gives a service ascribed to each one of these groups. If the difference between the probability that the service belongs to the first and latest group is greater than a threshold for trustworthiness, the agent being evaluated is considered a trusty agent.

There are two kinds of information that a witness can provide when it is queried about a target agent. If the target agent is one of its acquaintances it will return the information about it. If not, it will return referrals to the target agent that can be queried to obtain the information. These referrals, when queried, can provide the desired information or provide again new referrals. If the referral that finally gives the information is not far away to a depth limit in the chain, its information will be taken into account. The set of referral chains generated due to a query is a *TrustNet* similar to that used by Schillo et al. (Schillo et al., 2000) and in the Histos (Zacharia, 1999) model.

As we have said this model uses Dempster-Shafer theory of evidence as the underlying computational framework. In this case, to aggregate the information from different witnesses they use Dempster's rule of combination.

This model does not combine direct information with witness information (the two sources of information that takes into account). If direct information is available, that's the only source that is considered to determine the trust of the target agent. Only when direct information is not available the model appeals to witness information.

### 3.8. Sen and Sajja

In Sen and Sajja's (Sen and Sajja, 2002) reputation model, both types of direct experiences are considered: direct interaction and observed interaction. In the scenario where this model is used, observations are noisy, i.e., the observations may differ somewhat from the actual performance. Only direct interaction gives an exact perception of the reality. Reinforcement learning is the chosen mechanism to update the reputation value. Due to the noise underlying observations, the rule used to update the reputation value when there is a new direct interaction has a greater effect than the rule used to update the value when there is a new observation. The reputation value ranges from 0 to 1. A value greater than 0.5 represents a good performer and a value less than 0.5 represents a bad performer.

Agents can query other agents about the performance of a given partner. The answer is always a boolean value that says if the partner is good or not. In this model, liars are assumed to lie consistently, that means that every time they are queried, they return a good value for a bad target agent and vice versa. To decide, from the point of view of witness information, if a partner is good or not, the model uses the number of positive and negative answers received from witnesses. Knowing the number of witnesses and how many of them are liars, the model provides a mechanism to calculate how many agents should be

queried to be sure that the likelihood of selecting a good partner has at least a certain value. The subset of agents to be queried is selected randomly from the set of possible witnesses although the authors claim it is easy to add a smarter selection process based on a trust mechanism.

Because the objective of this work was to study how agents use word-of-mouth reputations to select one out of several partners, agents only use witness information to take decisions. Direct experiences are only used as pieces of information to be communicated to the others. Therefore, no indication is given by the authors about how to combine direct experiences with witness information to obtain a final reputation value.

### 3.9. AFRAS

The main characteristic of this model (Carbo et al., 2002) is the use of fuzzy sets to represent reputation values. Once a new fuzzy set that shows the degree of satisfaction of the latest interaction with a given partner is calculated, the old reputation value and the new satisfaction value are aggregated using a weighted aggregation. The weights of this aggregation are calculated from a single value that they call *remembrance* or *memory*. This factor allows the agent to give more importance to the latest interaction or to the old reputation value. The remembrance factor is modelled as a function of the similarity between (1) the previous reputation and the satisfaction of the last interaction and (2) the previous remembrance value. If the satisfaction of the last interaction and the reputation assigned to the partner are similar, the relevance of past experiences is increased. If the satisfaction of the last interaction and the reputation value are different, then it is the relevance of the last experience what is increased.

The notion of reliability of the reputation value is modelled through the fuzzy sets themselves. A wide fuzzy set for a reputation value represents a high degree of uncertainty over that value while a narrow fuzzy set implies a reliable value.

Recommendations from other agents are aggregated directly with the direct experiences. The weight given to each factor (old reputation value and new opinion) is dependent on the reputation that the recommender has. Recommendations coming from a recommender with a high reputation has the same degree of reliability as a direct experience. However, opinions from an agent with bad reputation are not taken into account. To calculate the reputation of recommenders, the agent compares the recommendation with the real behaviour of the recommended agent after the interaction and increases or decreases the reputation of the recommender accordingly.

### 3.10. CARTER ET AL.

The main idea behind the reputation model presented by Carter et al. (Carter et al., 2002) is that the reputation of an agent is based on the degree of fulfillment of roles ascribed to it by the society. If the society judges that they have met their roles, they are rewarded with a positive reputation, otherwise they are punished with a negative reputation.

Each society has its own set of roles. As such, the reputation ascribed as a result of these roles only makes sense in the context of that particular society. According to this, it is impossible to universalize the calculation of reputation.

The authors formalize the set of roles within an information-sharing society and propose methods to calculate the degree of satisfaction with each of these roles. An information-sharing society is a society of agents that attempt to exchange relevant information with each other in the hope of satisfying a user's request. They identify five roles:

- Social information provider: Users of the society should regularly contribute new knowledge about their friends to the society. This role exemplifies the degree of connectivity of an agent with its community. Each particular recommendation made by a user has a weight associated to it. This weight indicates the strength of the recommendation and is the product of a time decay factor and the reputation of the recommender. The degree to which the social information provider role is satisfied by a given user is calculated as the summation of all these weights, mapped in the interval [0,1].

- Interactivity role: Users are expected to regularly use the system. Without this participation the system becomes useless. The degree of satisfaction for this role is calculated as the number of user operations during a certain period of time divided by the total number of operations performed by all the users in the system during the same period.

- Content provider: Users should provide the society with knowledge objects that reflect their own areas of expertise. The degree of satisfaction is reflected by the quality of the information agents that belong to that user. The quality of an agent is measured considering how close is the subject of that information agent to the user's interest. The idea is that users that create information agents related to their areas of expertise will produce higher quality content related to their interest than those who do not.

- Administrative feedback role: Users are expected to provide feedback information on the quality of the system. These qualities

include easy-of-use, speed, stability, and quality of information. Users are said to satisfy this role by providing such information.

- Longevity role: Users should be encouraged to maintain a high reputation to promote the longevity of the system. The degree of satisfaction of this role is measured taking into account the average reputation of the user.

Given that, the user's overall reputation is calculated as a weighted aggregation of the degree of fulfillment of each role. The weights are entirely dependent on the specific society.

The reputation value for each agent is calculated by a centralized mechanism that monitors the system. Therefore, the reputation value of each user is a global measure shared by all the observers.

### 3.11. Castelfranchi and Falcone

The trust model proposed by Castelfranchi and Falcone (Castelfranchi and Falcone, 1998) is a clear example of a cognitive trust model. The basis of their model is the strong relation between trust and delegation. They claim that "trust is the mental background of delegation". In other words, the decision that takes an agent $x$ to delegate a task to agent $y$ is based on a specific set of beliefs and goals and this mental state is what we call "trust". Therefore, "only an agent with goals and beliefs can trust".

To build a mental state of trust, the basic beliefs that an agent needs are:

- Competence belief: the agent should believe that $y$ can actually do the task.

- Dependence belief: the agent believes that $y$ is necessary to perform the task or that it is better to rely on $y$ to do it.

- Disposition belief: not only is necessary that $y$ could do the task, but that it will actually do the task. In case of an intentional agent, the disposition belief must be articulated in and supported by two more beliefs:

  - Willingness belief: the agent believes that $y$ has decided and *intends* to do $\alpha$ (where $\alpha$ is the action that allows the goal $g$).
  - Persistence belief: the agent believes that $y$ is stable in its intentions of doing $\alpha$.

The first two beliefs compound what they call the *core trust* and together with the disposition belief, the *reliance*. Supported and implied by the previous beliefs, another belief arises:

- Fulfillment belief: if the agent "trust in $y$ for $g$", the agent decides: (i) not renouncing to goal $g$, (ii) not personally bringing it about, (iii) not searching for alternatives to $y$, and (iv) to pursue $g$ through $y$.

To summarize, trust is a set of mental attitudes characterizing the "delegating" agent's mind ($x$) which prefers another agent ($y$) doing the action. $y$ is a cognitive agent, so $x$ believes that $y$ *intends to do* the action and $y$ *will persist* in this.

### 3.12. ReGreT

ReGreT (Sabater and Sierra, 2001; Sabater and Sierra, 2002) is a modular trust and reputation system oriented to complex small/mid-size e-commerce environments where social relations among individuals play an important role. The system takes into account three different sources of information: direct experiences, information from third party agents and social structures.

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 graphs (sociograms) that define the agent social view of the world. These data bases feed the different modules of the system.

The *direct trust* module deals with direct experiences and how these experiences can contribute to the trust on third party agents. Together with the reputation model they are the basis to calculate trust.

The reputation model is divided in three specialized types of reputation depending on the information source that is used to calculate them:

- Witness reputation. If the reputation is calculated from the information coming from witnesses.
- Neighbourhood reputation. If the reputation is calculated using the information extracted from the social relations between partners
- System reputation. It the reputation value is based on roles and general properties.

The system incorporates a credibility module that allows the agent to measure the reliability of witnesses and their information. This module is extensively used in the calculation of *witness reputation.*

All these modules work together to offer a complete trust model based on direct knowledge and reputation. However, the modular approach in the design of the system allows the agent to decide which parts it wants to use. For instance, the agent can decide not to use *neighbourhood reputation* to calculate a reputation value or rely only on *direct trust* to calculate the trust on an agent without using the reputation module.

Another advantage of this modular approach is the adaptability that the system has to different degrees of knowledge. The system is operative even when the agent is a newcomer and it has an important lack of information. As long as the agent increases its knowledge about the other members of the community and its knowledge on the social relations between them, the system starts using other modules to improve the accuracy of the trust and reputation values. This allows the system to be used in a wide range of scenarios, from the most simple to the most complex. If the information is available, the system will use it.

In the ReGreT system, each trust and reputation value has an associated reliability measure. This measure tells the agent how confident the system is on that value according to how it has been calculated. Thanks to this measure, the agent can decide, for example, if it is sensible or not to use the trust and reputation values as part of the decision making mechanism.

The last element in the ReGreT system is the *ontological structure*. The authors consider that trust and reputation are not single and abstract concepts but rather multi-facet concepts. 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. For example, the reputation of being a good flying company summarizes the reputation of having good planes, the reputation of never losing luggage and the reputation of serving good food. In turn, the reputation of having good planes is a summary of the reputation of having a good maintenance service and the reputation of frequently renewing the fleet. Each individual can have a different *ontological structure* to combine trust and reputation values and a different way to weigh the importance of these values when they are combined.

## 4. Summary

In this section we show a table (see table I) that makes a summary of the models analyzed in this review from the point of view of the

classification dimensions presented in section 2. The abbreviations used in the table are the following:

| | | |
|---|---|---|
| Paradigm | N | Numerical |
| | C | Cognitive |
| Information sources | DI | Direct Interaction |
| | DO | Direct Observation |
| | WI | Witness Information |
| | SI | Sociological Information |
| | P | Prejudice |
| Visibility | S | Subjective |
| | G | Global |
| Model's Granularity | CD | Context Dependent |
| | NCD | Non Context Dependent |
| Agent behaviour assumptions | (see section 2.5) | |
| Model Type | Trust | Trust model |
| | Rep | Reputation model |
| General | × | No |
| | √ | Yes |
| | NA | Not applicable |

It is important to note that:

- We have described a set of classification aspects that allow a comparison between trust and reputation models. However, due to the diversity of such models, the classification aspects do not always fit exactly with the characteristics of the models and in some circumstances the classification for a specific model in one category or another is subjective according to our interpretation.

- We have considered only the features explicitly presented by the authors (without making suppositions on possible extensions).

- The decision of classifying the models as trust models or as reputation models is based on what the authors claim in their articles.

Table I. Comparison table.

| | Paradigm | Information sources | Visibility | Model's granularity | Agent behaviour assumptions | Boolean exchanged information? | Trust-Rep reliability measure? | Model type |
|---|---|---|---|---|---|---|---|---|
| S. Marsh | N | DI | S | CD | NA[5] | NA[5] | × | Trust |
| Online Rep. Models | N | WI | G | NCD | 0 | × | ×[8] | Rep |
| Sporas | N | WI | G | NCD | 0 | × | √ | Rep |
| Histos | N | DI + WI[7] | S | NCD | 0 | × | × | Rep |
| Schillo et al. | N | $\frac{DI}{DO}$, WI | S | NCD | 1 | √ | × | Trust |
| A.-Rahman and Hailes | N | DI, WI[1] | S | CD | 2 | 4 trust values | × | Trust Rep |
| Esfandiary and Chandrasekharan | N | $\frac{DI}{DO}$, WI, P | S | CD | 0 | × | × | Trust |
| Yu and Singh | N | DI, WI | S | NCD | 0 | × | × | Trust Rep |
| Sen and Sajja | N | $\frac{DI}{DO}$, WI[2] | S | NCD | 2[3] | √ | × | Rep |
| AFRAS | N | DI + WI[7] | S | NCD | 2 | × | √ | Rep |
| Carter et al. | N | WI[6] | G | NCD | 0 | × | × | Rep |
| Castelfranchi and Falcone | C | NA[4] | S | CD | NA[4] | × | NA[4] | Trust |
| ReGreT | N | DI + WI + SI + P[7] | S | CD | 2 | × | √ | Trust Rep |

(1) 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.
(2) Because the objective of this work was to study how agents use word-of-mouth reputations to select one of several partners, agents only use witness information to take decisions.
(3) Liars are assumed to lie consistently.
(4) In the description of the model it is not specified how the agents obtain the information to build their beliefs.
(5) There is no exchange of information between agents
(6) Besides information coming from other users (WI) there is a central authority that monitors the agents behaviour and uses that information to build reputation.
(7) The '+' symbol means the model combines the information sources to obtain a final trust/reputation value.
(8) The reliability is based on the number of ratings.

## 5. Discussion

As would be expected, the main sources of information 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, we think that a good mechanism to increase the efficiency of actual trust and reputation models (and also to overcome the lack of confidence in e-markets) is the introduction of sociological aspects as part of these models. It is true that in the actual e-markets this kind of sociological information is almost inexistent or it is not available to the participating agents. Therefore, nowadays, a model that uses this kind of information is not necessarily more useful than simpler models that only take into account (for example) direct experiences. It has no sense to increase the complexity of trust and reputation models if later on you have to use them in an environment where it is not possible to exploit their capabilities. Does it mean we have to give up doing sophisticated trust and reputation models? Certainly not. Electronic societies have to evolve to a new stage of complexity where interaction and links between their members become more relevant. This implies that a greater synergy between people working in the area of computational trust and reputation models and people dedicated to the research in the area of electronic institutions and norms has to be found.

Coming back again to table I, we see that only the ReGreT system, the AFRAS model and, in a way, the Histos model, propose methods to

combine different sources of information. The methods used by these three models to combine the information are a first step but, even in the case of the ReGreT system that has the most sophisticated method, they are far from being a general solution. They are too much dependent on the characteristics of the environment. A solution for this problem could be the use of non static methods, that is, adaptive methods that can modify how to combine the different sources of information according to the environment. This is not an easy task and we think it claims for a more deeper study.

Another aspect we have observed is that it is not usual to provide reliability measures of the calculated trust and reputation values, something that we think is very important.

Consensus is currently being reached on what trust is and reputation is in virtual societies. There are several works that help to give a precise and distinct meaning to both concepts (Conte and Paolucci, 2002; Mui et al., 2002; McKnight and Chervany, 1996). However, very few models propose links between both concepts. Our perspective is that reputation is one of the elements that helps to build trust on others. This relation between both concepts is something that should be studied with a great detail.

If we observe table I, it is clear that numerical modelling is the predominant paradigm used nowadays for the design of computational trust and reputation models. Possibly, the reason for that is the profile of people that is working in the area of multi-agent systems and e-commerce (economists and computer scientist) with a strong background in game theory and AI techniques. The question is whether this is the right approach. Numerical models have given good results in simple scenarios (simple with respect to the interaction complexity among individuals) like those currently present in Internet e-markets. However, when the complexity of the scenario increases, pure numerical models are not so good. These models reduce trust and reputation simply to a probability or perceived risk in decision makings (Castelfranchi and Tan, 2001). This seems to be too restrictive in scenarios where the complexity of the agents in terms of social relations and interaction is high.

A solution to this could be to explore other possibilities like, for instance, the cognitive approach and its combination with the numerical views. We think it is time to merge the tradition of sociologists and psychologists in the study of trust and reputation with the more pragmatic view that economists and computer scientists have explored.

Finally, analysing the models presented in this article we found that there is a complete absence of test-beds and frameworks to evaluate and compare the models under a set of representative and common

conditions. This situation is quite confusing, specially for the possible users of these trust and reputation models. It is thus urgent to define a set of test-beds that allow the research community to establish comparisons in a similar way to what happens in other areas (e.g. machine learning (UCI, 2003)).

## Acknowledgements

This work has been supported by the European project SLIE, IST-1999-10948, and the Spanish MCYT project e-INSTITUTOR, MCYT 2000-1414.

## References

Abdul-Rahman, A. and S. Hailes: 2000, 'Supporting Trust in Virtual Communities'. In: *Proceedings of the Hawaii's International Conference on Systems Sciences, Maui, Hawaii.*

Amazon: 2002, 'Amazon Auctions'. http://auctions.amazon.com.

Bromley, D. B.: 1993, *Reputation, Image and Impression Management.* John Wiley & Sons.

Buskens, V.: 1998, 'The Social Structure of Trust'. *Social Networks* (20), 265—298.

Carbo, J., J. Molina, and J. Davila: 2002, 'Comparing predictions of SPORAS vs. a Fuzzy Reputation Agent System'. In: *3rd International Conference on Fuzzy Sets and Fuzzy Systems, Interlaken.* pp. 147—153.

Carter, J., E. Bitting, and A. Ghorbani: 2002, 'Reputation Formalization for an Information-Sharing Multi-Agent Sytem'. *Computational Intelligence* 18(2), 515—534.

Castelfranchi, C. and R. Falcone: 1998, 'Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification'. In: *Proceedings of the International Conference on Multi-Agent Systems (ICMAS'98),Paris,France.* pp. 72—79.

Castelfranchi, C. and Y.-H. Tan: 2001, *Trust and Deception in Virtual Societies.* Kluwer Academic Publishers.

Celentani, M., D. Fudenberg, D. Levine, and W. Psendorfer: 1966, 'Maintaining a Reputation Against a Long-Lived Opponent'. *Econometrica* 64(3), 691—704.

Conte, R. and M. Paolucci: 2002, *Reputation in artificial societies: Social beliefs for social order.* Kluwer Academic Publishers.

Dellarocas, C.: 2003, 'The digitalization of Word-Of-Mouth: Promise and Challenges of Online Reputation Mechanisms'. *Management Science.*

eBay: 2002, 'eBay'. http://www.eBay.com.

Esfandiari, B. and S. Chandrasekharan: 2001, 'On How Agents Make friends: Mechanisms for Trust Acquisition'. In: *Proceedings of the Fourth Workshop on Deception, Fraud and Trust in Agent Societies, Montreal, Canada.* pp. 27—34.

Glickman, M. E.: 1999, 'Parameter estimation in large dynamic paired comparison experiments'. *Applied Statistics* (48), 377—394.

Grandison, T. and M. Sloman: 2000, 'A survey of trust in Internet application'.

Hume, D.: 1975, *A Treatise of Human Nature (1737).* Oxford: Clarendon Press.

Karlins, M. and H. I.Abelson: 1970, *Persuasion, how opinion and attitudes are changed.* Crosby Lockwood & Son.

Lashkari, Y., M. Metral, and P. Maes: 1994, 'Collaborative Interface Agents'. In: *Proceedings of the Twelfth National Conference on Artificial Intelligence,AAAI-Press.*

Marimon, R., J. Nicolini, and P. Teles: 2000, 'Competition and Reputation'. In: *Proceedings of the World Conference Econometric Society, Seattle.*

Marsh, S.: 1994, 'Formalising Trust as a Computational Concept'. Ph.D. thesis, Department of Mathematics and Computer Science, University of Stirling.

McKnight, D. H. and N. L. Chervany: 1996, 'The meanings of trust'. Technical report, University of Minnesota Management Information Systems Research Center.

McKnight, D. H. and N. L. Chervany: 2002, 'Notions of Reputation in Multi-Agent Systems: A Review'.

Montaner, M., B. Lopez, and J. de la Rosa: 2002, 'Developing turst in recommender agents'. In: *Proceedings of the first international joint conference on autonomous agents and multiagent systems (AAMAS-02), Bologna, Italy.* pp. 304—305.

Mui, L., M. Mohtashemi, and A. Halberstadt: 2002, 'Notions of Reputation in Multi-Agent Systems: A Review'.

OnSale: 2002, 'OnSale'. http://www.onsale.com.

Pearl, J.: 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann.

Plato: 1955, *The Republic (370BC).* Viking Press.

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

Sabater, J. and C. Sierra: 2002, 'Reputation and Social Network Analysis in Multi-Agent Systems'. In: *Proceedings of the first international joint conference on autonomous agents and multiagent systems (AAMAS-02), Bologna, Italy.* pp. 475—482.

Schillo, M., P. Funk, and M. Rovatsos: 2000, 'Using Trust for Detecting Deceitful Agents in Artificial Societites'. *Applied Artificial Intelligence* (Special Issue on Trust, Deception and Fraud in Agent Societies).

Scott, J.: 2000, *Social Network Analysis.* SAGE Publications.

Sen, S. and N. Sajja: 2002, 'Robustness of Reputation-based Trust: Booblean Case'. In: *Proceedings of the first international joint conference on autonomous agents and multiagent systems (AAMAS-02), Bologna, Italy.* pp. 288—293.

UCI: 2003, 'UCI Machine Learning Repository'. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Yu, B. and M. P. Singh: 2001, 'Towards a Probabilistic Model of Distributed Reputation Management'. In: *Proceedings of the Fourth Workshop on Deception, Fraud and Trust in Agent Societies, Montreal, Canada.* pp. 125—137.

Yu, B. and M. P. Singh: 2002a.

Yu, B. and M. P. Singh: 2002b, 'An Evidential Model of Distributed Reputation Management'. In: *Proceedings of the first international joint conference on autonomous agents and multiagent systems (AAMAS-02), Bologna, Italy.* pp. 294—301.

Zacharia, G.: 1999, 'Collaborative Reputation Mechanisms for Online Communities'. Master's thesis, Massachusetts Institute of Technology.