Lehrstuhl für
Wirtschaftsinformatik
Information Systems
Management

No. 46
October 2009

Bayreuther Arbeitspapiere zur Wirtschaftsinformatik

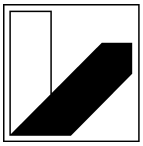
Tina Balke, Stefan König, Torsten Eymann

A Survey on Reputation Systems for Artifcial Societies

Bayreuth Reports on Information Systems Management

UNIVERSITÄT
BAYREUTH

ISSN 1864-9300



UNIVERSITÄT
BAYREUTH

Rechts- und Wirtschaftswissenschaftliche Fakultät
Lehrstuhl für Wirtschaftsinformatik (BWL VII)
Prof. Dr. Torsten Eymann

A Survey on Reputation Systems for Artificial Societies

Tina Balke, Stefan König, Torsten Eymann

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background and Motivation

The Internet has caused a revolution in trading. About a decade ago people had to sell the items they did not need anymore by means of a yard sale or an advertisement in the local newspaper, but nowadays they can offer their item on an auction site and potentially reach millions of interested people. Relatively cheap items that were not worthwhile advertising for in the past are now easy to sell on the Internet. As a consequence, sellers nowadays offer a wide range of products on the web, creating an abundance of choice for consumers. Before the Internet era it was virtually impossible to find very specific items such as particular chairs, books out of press, or carpets, just to name a few, but nowadays consumers have the opportunity to browse on different auction sites for the item they really want. Hence, both the ease with which consumers can offer an item to a wide audience, as well as the fact that consumers are more likely to find items that match their preferences, has caused trade in the Internet to grow exponentially [AS03]. Along with this success story however came the stories of people being victimized by fraudulent online sellers. These frauds cover a range from not delivering what has been promised, i.e. the overrating of a product's condition, to deliberate acts of theft and are a result of so-called asymmetric information.

Economist George Akerlof in his 1970 article [Ake70] explained the dangers emerging in markets where strong asymmetry of information and strong competition may deteriorate the quality of the goods exchanged, and eventually let the market disappear. The example of used cars retail is typical. People buy used cars that are advertised to be "in perfect conditions" at low prices. However most cars have hidden mechanical problems that become visible soon after the deal. Such cars are called "lemons" in the retailers' jargon. This can as well happen with artificial societies, where information asymmetry is intrinsic. Traditional protection authorities are often unable to trace back perpetrators and punish them. Very rarely contracts offer inspection of the merchandise before the payment. On the contrary, speed of dealing is one plus of the medium. Actually, countermeasures are being considered: resorting to brand name is one, but it is not effective when laymen are dealing with each other, as they may have no brand name or not a strong one. Important online auctions recommend using secure third party payment methods, such as Paypal. But again, this may become costly, and is no definitive solution as it relies simply on

another type of central authority enforcement. One solution to be further investigated is building trust by means of user-oriented rating mechanisms. That's why this paper analyses and compares existing trust and reputation mechanisms that were proposed to address this problem.

When looking at current literature, it has to be note that the concept of trust (and trust generation) is almost always mentioned in connection with the concept of reputation. The expansion of the topic on trust is rooted in the importance of trust for reputation concepts. Thus the fundamentals of reputation mechanisms are often derived from trust algorithms, and several papers presenting reputation approaches such as Zacharia's and Moukas' papers on Sporas and Histos [MZM99, Zac99, ZMM99], start by explaining images and trust-generating concepts and only as a second step analyse the distribution of evaluation information. Consequently, and furthermore for the reason of completeness, in this paper both trust and reputation concepts will be reviewed. Thereby the level of detail of the mechanisms varies from simple rough drafts to mathematical formalizations, which however leave some key questions concerning the implementation unanswered and finally to the description of algorithms with mathematical equations. However, in the context of this paper, it won't be possible to review all approaches in detail, as on the one hand, in the limited space of this paper no comprehensive all-embracing analysis is possible, and on the other hand, it would be necessary to first elaborate to what extent the implicated algorithms of some draft paper might be realizable. Nevertheless in order to get down to an executable but at the same reasonable description, the proposals shall be examined within the scope of several bigger categories.

So far several authors as for example [eRe06, JIB07, Kuh99, RJ96, SS05, YS00, ZMM99] have brought forward concepts for such categories, whereas the best known ones are the classification by Winter [Win99] which is based on the proposals of Rasmusson and Janson [Ras96] and Zacharia [ZMM99], the classification by Marti and Garcia-Molina [MGM06] who propose a taxonomy of trust and reputation systems in the P2P context and the classification by Sonnek and Weissman [SW05] who compare reputation systems in the Grid.

Looking at Winter's classification first (figure 1), it has to be realized that Winter distinguishes "soft"- and "hard control", where the first one refers to trust-based social mechanisms. Under "hard social control" Winter understands an agent-system-based social control that is not based on trust, but on the institutionalization of norms, such as resti-

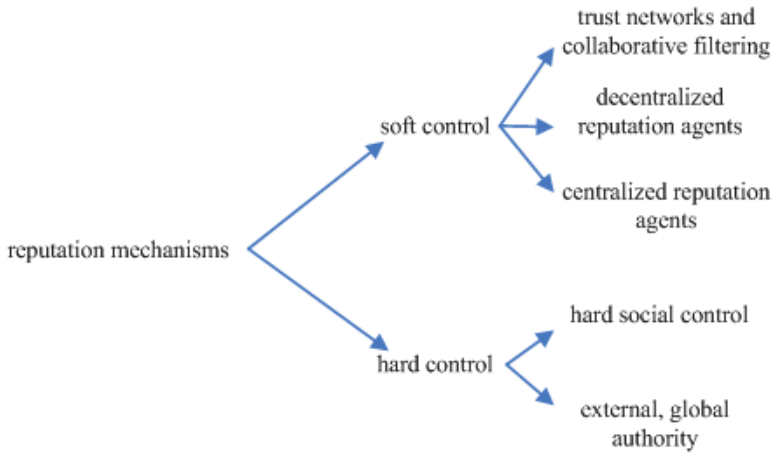


Figure 1: Classification of Reputation Mechanisms by Winter [Win99]

tution or expulsion options [Win99, p. 142 et. seq]. Putting it in other words, "hard social control" can be seen as the reputation of the system and its reputation mechanism. Hence, if agents trust that a reputation mechanism can filter out and penalize defecting agents, they are more likely to trade in that system. Although this classification approach is rampant, it has some disadvantages that make it unusable for this paper. Thus it is problematic to integrate cognitive concepts as well as the impact of 3rd-party-information in Winter's classification, however as it will be explained in chapter 1.2 these are integral parts of reputation mechanisms.

In contrast to Winter, Sonnek and Weissman include third party information in their model which is based on the works of Jøsang et al. [JIB07] and analyse the effectiveness reputation mechanisms in a service-oriented Grid context where clients request services from competing service providers [SW05]. They explicitly assume the possibility of intentionally given false reputation information, and by comparing several feedback algorithms regarding the inclusion of lying agents as well as mechanisms for identifying them, they derive an own feedback algorithm and prove its functionality by implementing and testing it for a specific scenario. However, looking at the overall classification approach, it has to realized that although Sonnek and Weissmann go into detail when analysing feedback mechanisms from several mechanisms, other classification aspects such as the storage of the data or the scoring and ranking of information is missing.

The latter aspect was taken up by Marti and Garcia-Molina. They identify three basic components of reputation systems that can be seen in figure 2, namely "information gathering", "scoring and ranking" and "response". Afterwards they break them down into separate mechanisms, categorize properties the mechanisms need to provide in order for

the reputation systems to fulfil its functions and discuss the implementation limitations and trade-offs that may prevent some of the properties from being met [MGM06].

Reputation Systems		
Information Gathering	Scoring and Ranking	Response
Identity Scheme	Good vs. Bad Behavior	Incentives
Info. Sources	Quantity vs. Quality	Punishment
Info. Aggregation	Time-dependence	
Stranger Policy	Selection Threshold	
	Peer Selection	

Figure 2: Classification of P2P Reputation Mechanisms by Marti and Garcia-Molina [MGM06]

The properties Marti and Garcia-Molina discuss are very detailed, however, due to their focus on P2P reputation systems, their discussion misses aspects about the logical storage of reputation information, which is decentralized in P2P systems, but can be centralize in artificial societies in general. Furthermore, although reasoning about their taxonomy in detail, Marti and Garcia-Molina do note analyse any existing reputation mechanism within their classification.

In this paper we try to overcome these drawbacks and develop a comprehensive trust and reputation mechanisms classification for artificial societies that will afterwards be used to analyse several existing mechanisms in chapter 3.

1.2 Image and Reputation

After this short introduction to the problem of reducing uncertainty and increasing trust in the artificial societies, in this section the terms trust and reputation shall be briefly explained. The definition used builds on social science and cognitive literature as within this area of research reputation and its effects have been discussed at length.

To start, we will define the term reputation as we understand it and relate it to the term *image* that will be of importance in the further course of the paper:

Image is a global or averaged evaluation of a given target on the part of an individual. It consists of a set of evaluative beliefs [MC00] about the characteristics of a target. These

evaluative beliefs concern the ability or possibility for the target to fulfil one or more of the evaluator's goals, e.g. to behave responsibly in an economic transaction. An image, basically, tells whether the target is "good" or "bad", or "not so bad" etc. with respect to a norm, a standard, a skill etc.

In contrast *reputation* is the process and the effect of transmission of a target image. The evaluation circulating as social reputation may concern a subset of the target's characteristics, e.g. its willingness to comply with socially accepted norms and customs. More precisely, we define reputation to consist of three distinct but interrelated objects: (1) a cognitive representation, or more precisely a believed evaluation (any number of agent in the group may have this belief as their own); (2) a population-level dynamic, i.e., a propagating believed evaluation; and (3) an objective emergent property at the agent level, i.e., what the agent is believed to be as a result of the circulation of the evaluation [CP02].

Putting it simple, an image is the picture an individual has gained about someone else (the target) based on his own previous interaction with that target. If using reputation, the individual expands the information source about the target beyond its own scope and includes the information of others about the target as well.

2 A 5-Stage-Process-Model for Reputation

In order to analyse and compare existing trust and reputation systems for artificial societies we propose a classification scheme which shall be briefly explained in this chapter. This classification scheme is firstly based on Sabater and Sierra [SS05], secondly on the ideas from the EU funded Project eRep [PEJ+09][1] and on the reputation process model by Padovan et al. [PSEP02] which was used for the highest layer roots.

In contrast to the classifications introduced in chapter 1.1, we do not see the reputation generation and usage dissectionable, but rather as a holistic process that stretches from the recording of transaction behaviour after one transaction to the usage of the reputational information for the next transaction. Therefore, our classification scheme is based on a five stages process, which takes place between two transactions of an agent, as demonstrated in figure 3.

[1] The ideas for classifying trust and reputation mechanisms were formulated formulated in Deliverable 1.1 [eRe06] of the eRep project. For more information see http://megatron.iiia.csic.es/eRep/?q=node/93.

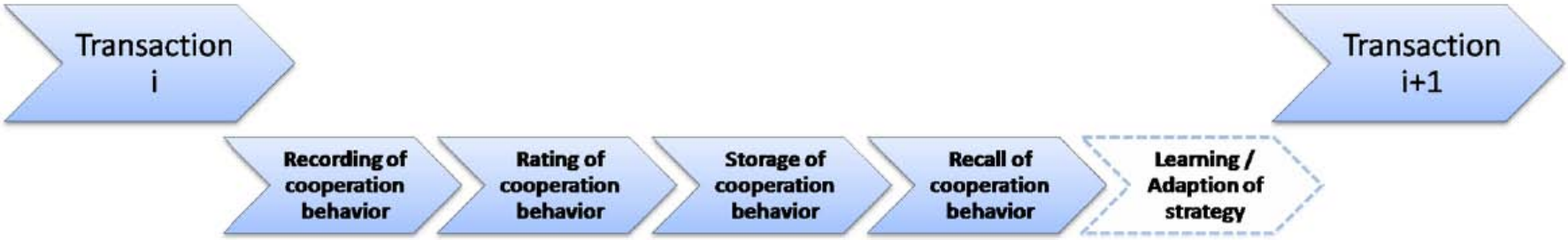


Figure 3: 5-Stage-Process Model
[PSEP02]

The stages start after the settlement of a transaction i and are called recording, rating and storage of cooperation behaviour, the recall of the former agent behaviour and last but not least the modification/adaption of agent's strategy. Along these stages, the trust and reputation models will be classified in this paper. Therefore, first of all the different stages will be explained in more detail.

2.1 Recording of cooperation behaviour

As a first step in the transaction phase model, after a settlement of a transaction, the cooperative behaviour of each transaction partner has to be recorded. Thus, all trust- and reputation models have to record the cooperative behaviour. These models are classified whether they are able to manage different contexts of cooperation or not.

> "If we trust a doctor when he is recommending a medicine it does not mean we have to trust her when she is suggesting a bottle of wine." [SS05]

This example makes clear, that trust and reputation are context dependent and as a consequence models need to be classified whether they are single- or multi-context models. A single-context trust and 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." [SS05]

Real multi-context models have to be differentiated from models, which seem to be able to handle multi contexts through instantiating more single-context models (each one for a special context).

> "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." [SS05]

Finally, it should be mentioned, that not in every application scenario a multi-context model is necessary. Adding the capability to deal with several contexts has to be paid with increasing complexity "and adds some side effects that are not always necessary or desirable" [SS05]. Furthermore, in special contexts, like eCommerce, it might be possible to put all trust and reputation information into one context without loosing too much of it.

2.2 Rating of cooperation behaviour

After the recording of the transaction partner's behaviour, the recorded behaviour (and ultimately the transaction partner) has to be rated [PSEP02]. The rating of cooperative behaviour phase looks at the recording and considers the algorithm used in the model to aggregate the rating values. Thereby, very broadly, three different approaches can be distinguished: Is there a cognitive approach, whose beliefs can not be aggregated mathematically, or is the model a mathematical one or even a composition of both approaches? In a cognitive approach mental states lead to trust other agents and perhaps lead to the decision to interact with these target agents afterwards. Game-theoretical/mathematical approaches consider trust and reputation as "subjective probabilities by which an individual A expects, that another individual B performs a given action on which its welfare depends" [Gam90]. Thus, trust and reputation are not, like in the former case, a result of a mental state, but a result of a pragmatic game with utility functions and a numerical aggregation of results in the past. [SS05] The result has to be mapped to a certain metric, which has to be unique within the system.

2.3 Storage of cooperation behaviour

After the rating of the transaction partner, the corresponding information has to be stored. Depending on the actual implementation, it can be stored by the transaction partners themselves or by a third party. Thereby, this categorization entry focuses on the logical data management and not on the physical data storage. The logical data management can be centralized or decentralized. To show by an example, that the logical view on the data is not necessary dependent from the physical data storage, one can think of a P2P network, which provides a logically centralized view of all data on the participating nodes. But the physical data storage is per se organized decentralized in such networks. In the

context of this work only the logical data management matters, because with a logical centralized view of all reputation data, the reputation model is able to use all data. With a decentralized view on reputation data, the agent and thus the reputation model is not able to see all these ratings. Instead, the individual agents have to make requests for other agent's former experiences with a certain target agent.

2.4 Recall of cooperation behaviour

Before entering the negotiation or agreeing upon a new transaction $i + 1$, the software agents recall available ratings about the prospective transaction partner (either based on their own experience or based on their own incomplete information in combination with the not necessarily trustworthy information provided by a third party) [PSEP02]. This dimension regards the information sources that the models take account of calculating trust and reputation values. The dimension is structured hierarchical. On the first layer it is differentiated whether there is an information exchange in the system. If there is no interaction in the system and only direct experiences or prejudices - that are more reliable however not available for all agents - are being used to assess the trustworthiness of a possible transaction partner, the systems are (usually) called Trust systems, as by definition reputation systems require an exchange of information in the system. In case information is exchange, further distinctions can be made:

Information exchange

If the trust- and reputation system provides information exchange between agents, there are two new dimensions to categorize the models. The first one is about the reliability of reputation values, the second one about the provision of semantic information. Witness and sociological information [SS05] are both two possible instances for exchanged information types.

- Meta-belief:
  Does the trust- and reputation model provide a measure of how reliable a specified trust or reputation value is? "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" [SS05].

- Value's semantic meaning:
  Third agents' evaluations depend on subjective cognition. To interpret this evalu-

ation, it is necessary to know the former evaluations of this evaluator in order to interpret the value correctly. For example a rating of 0.7 (in a continuous measure between 0 and 1) from an evaluator A for a target has to be interpreted different from the same evaluation value from evaluator B, if evaluator B is known for its low ratings.

- Type of exchanged information:
  The current model can be established in two big groups: those models which assume boolean information and those models which assume continuous measures. Although this seem to be a very simple difference, choosing one approach or the other, has a great influence in the design of the model. [eRe07] Models based on an aggregation mechanism usually have to use continuous measures.

Furthermore, the software agents can make use of the recalled information and rate them. Rating information influences the actual decisions of software agents in the selection of potential transaction partners. The most important impact criterion is the fact, whether the model allows cheaters or other malicious agents in the system and the agents to consider cheating and false information in their decision about cooperative behaviour.

In analogy to Sabater and Sierra we use three levels to show the degrees of agent' cheating:

- Level 0:
  The model does not consider cheating behaviour. There are many honest agents in the system, that the ratings provided by malicious agents are not preponderated.
- Level 1:
  The models classified here must be able to handle agents which hide information, but never lie.
- Level 2:
  The models classified into the highest level must even be able to handle lying agents and provide mechanisms to identify them.

2.5 Learning / Adaption of Strategy

Last but not least, the agents have to adapt their own future strategy to the experiences made before. The agent for example can utilize individual reputation thresholds to choose suitable interaction opponents. After a interaction has failed it might adapt this threshold

and increase it to choose more reliable partner in the future. On the other hand, if an interaction worked out well, the agent might be willing to decrease this threshold in order to select a transaction partner with a higher matching probability when for examples markets are used match supply and demand.

Nevertheless, this issue is very important in the field of trust and reputation model almost all trust and reputation models we will consider in section 3 exclude proposals on this topic due to the high context-dependency of the strategy.

The following section will consider different reputation mechanisms following the special view defined in this section.

3 Reputation Mechanisms for Artificial Societies

3.1 Marsh

The first concept that shall be analysed with the help of the 5-stage-process-model is the *Trust*-concept by Stephen Marsh. [Mar94] Looking at the recording of the cooperation behaviour, it has to be noted that *Trust* is a multi-context model (i.e. it can to handle different contexts of cooperation) as Marsh distinguishes between three types of trust in his thesis: basic trust, general trust and situational trust. [Mar94, p. 55]

Basic trust is the trust the agent has independently from the current transaction-offering agent. It is calculated from all the experiences accumulated by the agent. *General trust* is the trust an agent has in another agent without taking any specific situation or context into account. This is done in the *situational trust* calculation, in which Marsh uses the formula

$$T_x(y, \alpha) = U_x(\alpha) \times I_x(\alpha) \times \widehat{T_x(y)} \tag{1}$$

to determine the situational trust $T_x(y, \alpha)$ of agent x (the truster) in agent y (the trustee) in the specific situation α[2], resulting of the subjective utility $U(\alpha)$ and the subjective importance $I(\alpha)$ of the situation. [Mar94, p. 62] Thereby the utility U_α can take values of the interval $[-1, 1)$, whereas the importance of the situation comes from $[0, 1]$. The last term of the equation $\widehat{T_x(y)}$ refers to a possible weighted mean value over the evaluations of the general trust for agent y by agent x in the past $(\widehat{T_x(y)^{t-n}}; \ ...; \ \widehat{T_x(y)^{t-1}})$. As the trust function itself, it can take values from the interval $[0, 1)$ [3]. Along the lines of situational trust, Marsh furthermore makes intensive calculations depending on the standard of knowledge an agent x perceives an agent y to have for a situation α and specifies equations for three states of competence he distinguishes. These three states are [Mar94, p. 73 et seq.]:

1. The state in which agent is not known in the specific or a similar situation (e.g. if the agent has never trade with the other agent before).

2. The state in which an agent is known but not in the specific or in a similar situation (e.g. in case the agent has traded with the other agent before, but in a totally

[2] When talking about *situations* Marsh refers to the comparability of actions. According to Marsh trustworthiness always refers to experiences in comparable situations.

[3] Marsh excluded the value 1, as it would represent blind trust, and consequently would not be expedient [Mar94, p. 57].

different context).

3. The state in which the agent is known and trusted in the specific or a similar situation (e.g. in case the agent has traded with the other agent before, in the same context).[4]

The perceived-competence-variables, Marsh calculated for the three cases, are then used in the rating process when evaluating the cooperation behaviour. In this process, Marsh discusses several "dispositions", namely optimistic, pessimistic and realistic agents, and calculates their trust evaluations from the past up to a certain point that is either determined by the agents' memory size or the number of previous encounters. Hence, for the estimation of its partner's trustworthiness, an optimistic agent takes into account the maximum value of all previous comparable transactions, whereas a pessimistic one will consider the minimal value.[5] Although Marsh included these cognitive "dispositions" his concept analyses trust purely mathematically, as not mental states lead to the decision to trust another agent, but subjective probabilities.

Once, Marsh has calculated the values for the situational trust, they are stored locally by each agent to be then used in case an agent is offered a transaction with the evaluated agent once more. Hence, the system Marsh envisaged decentralized in terms of the data management [Mar94, p. 5]. If an agent is approached with a transaction offer, Marsh calculates a *cooperation threshold* to determine whether the agent should interact with the potential transaction partner, or not:

$$Cooperate\ \ Threshold_x(\alpha) = \frac{Perceived\ \ Risk_x(\alpha)}{Perceived\ \ Competence_x(y,\alpha) + \widehat{T_x(y)}} \times I_x(\alpha) \qquad (2)$$

Thus in case, the *Cooperation Threshold* is reached or passed, a transaction can take place. In case several agents at the same time offer a transaction and more than one extravagates the threshold, Marsh discusses different procedures. Thus he points out that the agent has the option either to choose the generally most trustworthy offering agent, or the one most trustworthy with regard to the situation, etc. It even might make its decision depending on all procedures and choose the agent which does best in most

[4]Thereby, the equations Marsh specifies for case number (2) and (3) are expansions of the standard formula for the perceived competence $Perceived\ Competence_x = T_x I_x(\alpha)$, that is used in the first case. Concerning the equations themselves, it has to be noted, that although specifying equations, Marsh himself tends to use estimations instead of exact calculations.

[5]For the realistic or pragmatic agents, Marsh consults the mean value of all previous encounters that from the situation's point of view are comparable to the new one.

of them. As these calculations show, Marsh's *Trust*-concept does not take into account any third-party-information and as such does not calculate any reputation-values. Hence, when looking at the recall phase, *Trust*, as its name already indicates, has to be identified as a trust system.

If, finally, the agents decided to trade with one another, after each transaction the general trustworthiness in terms of the *Cooperation Threshold* is adjusted by each agent using multiplication. [Mar94, p. 79 et seq.]:

Marsh showed this mechanism using an iterated prisoner dilemma which is exemplarily shown from agent x's perspective in the following table:

	Y *cooperates*	Y *deceives*
X *cooperates*	$T_x^{t+1} = T_x^t \times 1,01$ $T_x(Y)^{t+1} = T_x(Y)^t \times 1,10$	$T_x^{t+1} = T_x^t \times 0,99$ $T_x(Y)^{t+1} = T_x(Y)^t \times 0,90$
X *deceives*	$T_x^{t+1} = T_x^t \times 1,05$ $T_x^{t+1} = T_x^t \times 1,01$	$T_x^{t+1} = T_x^t \times 0,95$ $T_x^{t+1} = T_x^t \times 0,90$

Table 1: Adjustment of agent X's *Cooperation Threshold* for Y

In the table it is important to note that the number values can be chosen freely by each agent, although Marsh strongly suggested to penalize deceitful behaviour more strongly than to award the cooperative one, as in the real world it is easier to lose, than to gain trust.

When trying to evaluate Marsh's approach, all in all it does not seem very complex but easy to implement. However, it harbours several problems. Thus, when commenting on Marsh's approach Schillo [SFR00, p. 36 et seqq.] for example, points out effects that are owned to math. Marsh, at least partially, sees these problematic cases as well; however he sticks to the conviction that his formalization models trust successfully. [Mar94, p. 143] The problems addressed by Schillo can be exemplified when looking at the equation for the situation trust (see equation 1).

In this equation Marsh uses a product whose factors can be negative. This leads to the unusual side effect that the situational trust is positive in case the agent does not have any trust in his potential trading partner and hence attributes a negative trust value to

him and at the same time judges the utility of the situation negative as well.[6] Marsh describes this effect as "machiavellian" and reckons that to a certain extent it can be useful to leave transactions that do not seem useful or important to these agents who are less trustworthy. Further equivocal effects that result from the use of multiplication and the interval [-1,1]lead to the problem that in the case of trust- or utility-indifference (i.e. $\widehat{T_x(y)} = 0$ or $U_x(\alpha) = 0$) by the agent the situational trust is assigned 0, which is a value that is difficult to interpret.[7] When these mainly mathematical problems are put aside, several further questions remain for the discussion of Marsh's concept. Thus one must analyse whether the proposed formalizations would result in a reliable complete system if they were implemented in artificial societies, for instance. For answering this question it seems reasonable to examine Marsh's approach for its functionality. Thereby especially the aspect of social control (i.e. to what extent deceiving agents can be filtered out and excluded from the market-place) is of special interest. [Mar94, p. 138]

Generally, the discussed formalization seems to be suitable for reducing fraud. Thus, if agent y cheated agent x, there is a high probability that in the future agent x will not trade with y again. However, the values for modification of the general trust have to be adjusted drastically to reduce the general trust in case of fraud. With these adjustments at least pessimistic and realistic agents can reasonably protect themselves against deception. Nevertheless, as no communication between the agents exists, a streetwise agent can systematically betray all other agents once without the system stopping it. In this respect, the request for a secure total system cannot be fulfilled by Marsh's mechanism.[8] Besides the question of functionality, it has furthermore to be asked whether the underlying sociological assumptions endure the limitations of a technical system. For his definition of trust, Marsh chose the work of Morton Deutsch, Niklas Luhmann, Bernard Barber and Diego Gambetta as a starting point, whereas the latter had the biggest impact on him. [Mar94, p. 25 et seqq.] This influence can, for example, be seen in the adoption of situation- and agent-specific cooperation thresholds as well as in the conception of situational trust as a product of utility and importance on the one hand and trustworthiness on the other. Thus it can positively be noted that through the inclusion of the situation

[6] Similar arguments can be found against the *Cooperate Threshold* formula, if for example the $Perceived\ Competence_x(y, \alpha) + \widehat{T_x(y)} = 0$.

[7] In his comments on "No Trust and Distrust" Marsh himself distinguishes among four possible interpretations (relating to general trust) of the value 0.

[8] As a matter of fairness, it has to be said that the establishment of a totally secure system was not the departure point for Marsh's considerations. He concentrated on the establishment of trust relations and their implications for co-operation, instead of the security of the system and even proposed himself to establish a communication between agents as possible solution. [Mar94, p. 116]

and its significance, Gambetta's idea of social arrangements influencing the necessity of trust [Gam90, p. 220] was taken into consideration. Furthermore Marsh considered the unequal speed in the growth and decline of trust respectively. Putting it in a nutshell, "Trust" represents an approach that is fairly easy to implement. It is based on the idea that agents calculate trust-estimations independently from other agents based on their own experiences. The practicability of the mechanism is, as a result, dependent on the adjustment of several parameters (e.g. the decline of trust in case of fraud or the memory margin). As explained before, the usage of the interval [-1,1], the quantitative implementation and the negligence of third-party information seem questionable and present challenges.

To sum up the findings about Marsh's *Trust*-concept, it has to be said that it has been formative for many further developments and algorithms in the artificial intelligence trust and reputation research. However, besides the question of whether his estimation of the cooperation threshold and the trustworthiness are sustainable, one of the main points of criticism in Marsh's model is that in the rating phase his agents only rely on their own observations and do not include any third-party information in the calculation of the trustworthiness.

Some of the authors who tried to propose models tackling this problem were Alfarez Abdul-Rahman and Stephen Hailes [ARH97a, ARH97b], Lars Rasmusson and Sverker Jansson [Ras96, RJ96], Michael Schillo et al. [Sch99, SFR00] as well as Bin Yu and Munindar P. Singh [YS00]. Although these approaches do have the same basic idea in common - namely, that the experiences of other agents in the network can be included when calculating trust and reputation values and when searching for the right transaction partners - they vary when it comes to the questions of how to weight the third-party information and how to deal with the friends of friends as well as information from agents who seem to be not very trustworthy. One of the best known approaches in this category is the *TrustNet*-concept of Michael Schillo [SFR00] which shall now be reviewed in more detail.

3.2 Schillo

The single-context *TrustNet* approach by Schillo et al. was published in 1999 as an expansion of a "Trust-concept" by Castelfranchi et al. [CdRF97]. Originally intended for implementing trust in MAS, compared to Marsh the concept goes one step further in

the recall phase and furthermore allows for agents to exchange witness information with one another and to use this information for their cooperation decisions. Thereby Schillo considers not only the pure gossip, but includes the lack of validity of the benevolence assumption in open systems by including assumptions about the honesty and the degree of altruism of other agents in his thoughts (level 2 in the modification phase). However, in total, only a single trust value is attributed to each agent, hence in the recording phase no context-dependent differentiation of the trust values takes place.

The basic idea and first central element of Schillo's concept is based on a repeated 5-step-Prisoners Dilemma that works as follows:

(1) in the first part of a round the agents have to pay a participation fee. (2) The second step is negotiations in which the agents can make (false) announcements about their intentions. (3) A PD-round is played and both agents publish their moves simultaneously. (4) The results of the round are announced. They can be seen by the two agents involved and by the agents in their neighbourhood. (5) The fees are paid out.

Schillo's idea now is that in the phase (2) the agents have the opportunity to interview agents they know about the unknown trading partner. Furthermore the results from phase (4) can be used to update the *TrustNet*. As a result, agents who did not keep their promises from phase (2) can be spotted relatively fast not only by their trading partners, but by the neighbourhood as well. Hence Schillo concludes that untrustworthy agents won't be able to find partners after some rounds and thus cannot earn any more points [SFR00, 829 et seqq]. How this mechanism for estimating the trustworthiness works in detail shall now be explained.

Schillo's model is based on the probability theory and therefore has to be classified as mathematical model in the second (i.e. the rating) stage of the process model, although he considers cognitive concepts such as altruism[9], etc. in the model. In order to be able to assess another agent's trustworthiness, an agent generates models about the agent's honesty and altruism. Thereby an agent Q's "honesty" $E(Q)$[10], which is supposed to express the probability that an interaction is processed according to the announcements, is represented by the coefficient of the number of transactions that was processed according to the announcements against the total number of transactions. Along this lines, for

[9]The agent acts altruistically if it cooperates irrespective of his opponent.

[10]The letter E is derived from the German "Ehrlichkeit" that can be translated with honesty, fidelity or forthrightness in English.

simplicity reasons, Schillo models agent X perception of the honesty $E_x(Q)$ of agent Q, and about Q's altruism affinity $A_x(Q)$[11] respectively. [Sch99, 49 et seqq.]

When it comes to the third stage in the process model - the storage of cooperation behaviour stage - the second central element of Schillo's approach comes into play: the *TrustNet*, which has stored all information centrally from the logical data management point of view and therefore serves as memory for other agents' trust estimations. The *TrustNet* itself is a data structure and has the form of connected vectors as shown in figure 4:

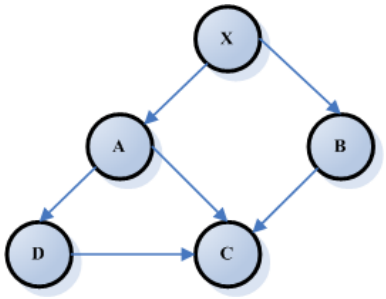


Figure 4: Schillo's Trust Net
[Sch99, p. 73]

In the model the nodes represent the agents and the edges the observations. Observations by the agents are sets of triples of the form $(PD-round, alt, ehrl)$, whereas *alt* (altruism) and *ehrl* (honesty) can either take the boolean values yes or no, or can be kept secret [Sch99, p. 52]. Each agent has got his own data structure in which it is represented by the root-node X (no incoming - and only outgoing edges), while the other nodes represent the neighbour-agents which can give direct (direct connection of root node and agent-node) or indirect (witness) information about the trustworthiness of a potential trading partner. The information is stored with the edges and the derived assessments are filed as a model of honesty and a model of altruism by the respective nodes [Sch99, 72 et seqq.]. In general this graph may have cycles (in figure 4 this is the case when looking at $A \longrightarrow D \longrightarrow C$), however with the help of an adequate algorithm, Schillo tries to eliminate them as far as possible in order to have only the edge with the highest information content remaining [Sch99, p. 79]. Besides this difficulty, the integration of multiple witness information about an agent poses a further problem in Schillo's concept. That is why he introduced an additional recursive algorithm which, by dint of probabilistic considerations about the motivation of the agent[12], delivers an integrated estimation of the honesty and altruism

[11] As Schillo uses probability value, the values all lie in the interval $[0, 1]$.
[12] In this scenario Schillo assumes that agents want to cast a damming light on other agents and thus he

of another agent.

But how does an agent calculate the trustworthiness of his potential trading partner in the recall stage of the transaction process? Starting from the generated honesty and altruism models the trustworthiness V[13] of an agent Q which offered a transaction to agent X can finally be calculated, where the V-value represents the probability that the cooperation will be successful [Sch99, p. 78]:

$$V_x(Q) = \frac{A_x(Q)}{A_x(Q) + (1 - A_x(Q))(1 - E_x(Q))} \tag{3}$$

When comparing Schillo's model to Marsh's, it is unmistakable that Schillo attaches more importance to the derivation and justification of the formulas and equations used by him. Furthermore, in contrast to Marsh's concept, no problems concerning the multiplication of negative values appear, as Schillo uses probability values for describing certain behaviours. In addition he predicates social norms as the equivalent to trustworthiness and thus uses honesty and altruism as well as the possibility of the deliberate concealment of information in his model. Comparing Schillo to Marsh, Schillo's definition of trustworthiness $V_x(Q)$ has almost the same meaning as the situational trust $T_x(Q, \alpha)$ in Marsh's model and is used to select trustworthy cooperation partners by consulting and evaluating witness information. Instead of a "simple" memory for situational trust values, Schillo proposes a data-structure, the *TrustNet*, in which trust information for the agents in the net and the gossip as well as the altruism and honesty values are stored. As mentioned before (when referring to the recording of cooperative behaviour-stage) Schillo does not, however, include any situational observations as they would make the concept far more complex. Thus no equivalent to Marsh's situationally dependent basic trust of an agent is present in the *TrustNet*.

However Schillo's attempt poses problems as well. For example, he does not go into detail about the problematic situation of new agents who are completely unknown to the *TrustNet*. He only mentions that in phase (2) of his model all agents are informed about the self-description of the new agent and can cooperate with it or not. The new agent can then choose the most trustworthy one from the ones who are willing to transact with him and a deal is agreed on. Concerning the functionality of the whole system (to increase the

concludes that they will be more willing to give evidence about cheating than about honest behaviour. Based on these assumptions a stochastic-method-based heuristic can be derived that helps to determine the number of positive behaviour observations an agent kept secret.

[13] The letter V is derived from the German "Vertrauenswürdigkeit".

preparedness for cooperation and to reduce defective behaviour), Schillo noted that in the long run, if enough models for the agents have been developed, altruistic agents perform better than their less friendly counterparts. Hence, Schillo's approach at first glance seems to create incentives to be honest and cooperative. However it is unclear whether these qualities can be found in artificial societies as well, as they can do have an open structure and agents can leave and enter the market as they wish. Furthermore, in the case of cooperative behaviour, "cooperative cliques" may form that impede new agents from breaking into the market and just leave them with the possibility of bargaining with egoistic agents who are not members of the 'cooperative cliques'. As a result the question can be posed whether personal and public knowledge of the agents is sufficient, or whether additional variables such as the information bundles from Sztompka [Szt99] should be included in the calculation of trustworthiness.

Summing up the information about the *TrustNet* model, it has to be acknowledged that both its theoretical foundation as well as its implementation are based on Schillo's ideas of an openly played PD with a partner selection. Hence, the question comes up as to what extent the demand to apply the model to other areas can be fulfilled. Schillo sees a direct connection between the PD and the situation in artificial societies and talks of a direct interrelation of both. Nevertheless he makes little comment about the actual implementation in artificial societies. Thus he does not refer to the possible situation of a complete information asymmetry (some agents trade on the market for a long time and some completely new agents enter the market) and he does not account for the scalability of his model and whether it will work on large markets with potentially 1000 up to 100,000 agents. This last problem, the lack of scalability, becomes less important when considering that in large markets possibly only small cliques of agents who know one another emerge. However this results in the question, as to how the dilemma of neighbourhood and publicity shall be solved in open markets. Furthermore the question of the data storage has to be solved. Thus Schillo specifies the storage complexity of each agent with $O(n^2r)$ whereas n indicates the number of familiar other agents to an agent and r accounts for the number of played (and stored) PD-rounds. Hence the storage complexity is higher than in Mash's concept ($O(nr)$), as in the *TrustNet* model the information from the edges has to be stored additionally. Concerning the time complexity, *TrustNet* uses the principle of lazy evaluations; viz recalculations of the evaluations are executed in case a new palpable offer is at hand. Thereby inclusion of edges takes place in the cubic time ($O(n^3)$) and consequently is relatively costly.

To recap, *TrustNet* is an interesting approach for calculating trust and reputation. It is more complex than the model of Marsh; however, both the nature of the trustworthiness assessment as well as the inclusion of inter-agent communication make the concept more substantiated. Nevertheless it remains questionable whether the desired functionality (to increase the preparedness for cooperation and to reduce defective behavior) can be achieved outside the controlled and synchronized test scenario.

3.3 Rasmusson and Janson

Another scientist who studied reputation mechanisms for artificial societies and thereby focused on the inclusion of third party information in the recall-stage was Lars Rasmusson [Ras96] who worked together with Sverker Janson [RJ96, RRJ97] on general concepts to ensure security in open agent-based networks and tried to prove them with the help of simple simulation systems. As a result of their work, Rasmusson and Janson, mainly for scalability and monopoly avoidance reasons, plead for a decentralised organized reactive security approach. Hence, according to their ideas, security shall primarily originate from agent behaviour and shall not (or only additionally) be imparted by central institutions. For the realization of their approach, Rasmusson and Janson sketched several ideas that all used a centralized approach in the storage phase. These ideas included the implementation of Trusted Third Parties (TTP) for financial transactions or the introduction of special reviewer agents for example, but, most importantly, they made proposals for a social approach: thus, agents should be in the position to actively obtain information about other agents and to pass on heard gossip. However, a problem occurs in this regard, which was already discussed by Schillo: at least in highly competitive environments, agents have a certain incentive to lie, as negative information about an agent can result in competitive advantages for its competitors [RJ96, p. 14]. Therefore possible amplification of gossiping can lead to agents advertising themselves. A possible solution to the problem of the deliberate transmission of false information that was proposed by Rasmusson is the use of money, so that agent A which wants to praise itself or another agent (B), has to pay agent C to remember him or B [Ras96, p. 25]. Therefore, Rasmusson's and Janson's mechanism can be said to be a kind of recommendation mechanism, where the recording of the behaviour (in stage one of the 5-stage-process model) is not context dependent and only takes place if the recommended agent (which can be the recommender itself) pays for it. The rating the the recorded behaviour was than aggregated mathematically.

Besides these social proposals, as Ackerlof [Ake70] did, Rasmusson wanted to demonstrate the importance if reputation mechanisms for market stability and market success and in this context achieved simulation results that are of further interest, although he only included favourite choice approaches in his studies [Ras96, p. 29 et seqq]. His results indicated that a complete orientation to price (without looking at any trust or reputation values) reduced the total quality of the market and stable conditions were only achieved after several buying-agents had declared bankruptcy. In the case where the price-orientation was combined with the restriction of a single market place a monopoly developed. However, as soon as he included favourite choice approaches in his experiments and consequently allowed for the agents to remember the behaviour of their trading partners, the proportion of defecting agents decreased and (in combination with a multi-market concept) a stable and qualitatively high-value situation was achieved. Ramusson assumed that these positive effects might even be strengthened through the implementation of gossip; however, he did not prove this idea in his experiments.

3.4 Abdul-Rahman and Hailes

A further reputation concept that shall be discussed is one of Alfarez Abduhl-Rahmen and Stephan Hales [ARH97a, ARH97b, ARH00]. Their model that pursues a perspective different from Schillo, is directly related to internet-based MAS and is supposed to help to implement trust as the basis of informal-, short-term-, or commercial ad-hoc transactions. Therefore they propose that every agent carries along a network of trust relationships in a database, hence hence information are stored decentralized. Abduhl-Rahman and Hailes define a 'trust-relationship' as a vectored connection between exactly two entities, which in some circumstances can be transitive. In this way they distinguish between direct trust relationships ("Alice trusts Bob.") and recommender trust relationships ("Alice-trust-Bob recommendations about the trustworthiness of other agents"). Consequently, Abdul-Rahman and Hailes make a context-dependent distinction in the recording phase. Thus, an interesting contrast to the other formalizations lies in the fact that due to the qualitative nature of trust, Abduhl-Rahman and Hailes do not work with probability values or the $[-1, 1]$ interval, but use a multi-context recording model, i.e. discrete values that are not universally valid, but are related to certain trust categories ("Alice trusts Bob, concerning "table"-transactions. However, she does not trust him when it comes to "chair"-transactions."). The discrete values used can be seen in table 2.

Value	Significance for direct trust relationship	Significance for recommender trust relationship
-1	Distrust - completely untrustworthy	Distrust - completely untrustworthy
0	Ignorance - cannot make trust-related judgement about entity	Ignorance - cannot make trust-related judgement about entity
1	Minimal - lowest possible trust	
2	Average - mean trustworthiness (most entities have this trust level)	
3	Good - more trustworthy than most entities	
4	Complete - completely trust this entity	

Table 2: Discrete trust value after [ARH97b, p. 53]

As a result, Abduhl-Rahman and Hailes define reputation as a 'troika' ($agent - ID$, $Trust - Category$, $Trust - Value$). Each agent stores such reputation information in his own data-base and uses it to articulate recommendations. However, in their concept Abduhl-Rahman and Hailes make no comments on how agents derive these evaluations.

Nevertheless, the core of Abduhl-Rahman and Hailes' papers allegorize their thoughts about a recommendation protocol that can be used to communicate recommendation requests and statements as well as updating inquiries within the MAS. In the protocol, a recommendation request, for example, is passed on until one or more agents are found which can give information for the requested category and which is trusted by the penultimate agent in the chain. Based on this idea Abduhl-Rahman and Hailes propose a mathematical algorithm for the rating phase in which the requesting agent can use the following equation to calculate the trustworthiness of a recommendation. For $tv(R_x)$ as the recommender trust value of the different recommendations of the involved agents and $rtv(T)$ as the trust value articulated by the last agent[14] the trustworthiness result from the following equation:

$$tv_r(T) = \frac{tv(R_1)}{4} * \frac{tv(R_2)}{4} * ... * \frac{tv(R_n)}{4} * rtv(T) \tag{4}$$

[14]In case an agent receives more than one recommendation about another agent, the values are averaged.

In summary, both the missing justification for the special way of calculating reputation (that was admitted by the authors) and the absence of any information on how agents derive their direct trust values, the lack of inclusion of any cheating possibility (in the modification phase) as well as the necessity for a global standardized category-ontology indicate serious conceptual problems for the model. Nevertheless, due to its qualitative and at the same time algorithmic approach, the outline seems worth mentioning, whereas especially the idea of conceiving trust and distrust (referring to a certain category) not as a continuum but as a condition has to be highlighted.

3.5 Regan and Cohen

Regan and Cohen [RC05] proposed a system for computation of "indirect" and "direct" reputation in a computer-mediated market. Recognizing that centralized public values of reciprocal ratings between agents may bring about collusion, blackmailing and retaliation, the authors propose that only evaluations about sellers should be taken into consideration. Their assumption is that reputation is to be used to find quality sellers rather than buyers, as the former have more control over exchanges, especially in an A2A (all to all) online market. Evaluation from buyers on sellers should then be transmitted opaquely with respect to the latter. Their objective is simply to propose a mechanism which reduces the "undesirable practices" possible in actual reputational online applications, especially on the part of sellers, and thus to prevent the market from degenerating into a "lemons market" [Ake70] where only low-quality goods are listed for sale. For calculating "direct reputation" the authors refer to a mathematical rating model proposed by Cohen and Tran [TC01]. This proposal postulates the partition of the seller pool into three sets: (1) the ones with a good image, (2) the ones with a bad image, (3) the ones with an uncertain image in the eyes of the prospective buyer. Reputation is then an integer from $[+1, -1.0]$ respectively where the zero value is assigned to unknown sellers, $+1$ to the "good" ones and -1 to the bad ones. To calculated the reputation value, the seller's contract performance is confronted with the buyer's satisfaction threshold. This may be set as to make it difficult to gain a good reputation, and relatively easy to lose it. Furthermore the transaction's value is taken into consideration what makes Regan and Cohen's model a multi-contextual one in terms of money as Regan and Cohen distinguish between expensive and inexpensive transactions and do not treat all transactions equally.

This model has integrated introducing agents into the role of "advisors", that is, agents

owning direct information about the target. Advisors own a reputation as informers. To model this request process, the authors propose two examples: (1) peer-to-peer networks like Gnutella [Rip01], where a network's information list is available to every node through a software-connecting all peers, which is used to send requests and receive information; and (2) some centralized server maintaining every evaluation from agents to which requests are to be sent. Some corrections should be applied to account for possible adviser's bias. First, only highly reputed agents are being asked. Secondly, advisers with similar preference patterns as one's own are to be preferred. Information provided is not being weighed on their reputation (as it happens in Sporas, see chapter 3.6). Information with high deviation from the mean is ruled out. Following the actual transaction, the seller's as well as advisers' reputations are updated. The authors admit that their model alone has a problem with agents abandoning pseudonyms with a bad reputation, because it assigns a lower value to an agent with a bad reputation than to an unknown entrant or "newcomer". Among the possible solutions mentioned there is: a third-party authority assigning not renewable unique identifiers to agents; monetary incentives for keeping the same pseudonym over time, with some form of market entrance cost; the possibility of selling a good reputation when deciding to leave the market; some cautionary sum to be deposited on entrance that would be given back when leaving with a neutral-to-good reputation.

This model presents many drawbacks: its modelling of reputation transmission is unsatisfactory, as it does not consider the possibility of spontaneously giving reputation information to some selected recipient. Here, like elsewhere, only asking and answering is contemplated. However, despite its evident simplicity, this model tries to tackle the problem of collusion between rating agents, positioning the rating activity in a way that only sellers are evaluated in an opaque way with respect to them. This is obviously only part of the solution, because it would be good to have some indicator of the buyer's reputation as well for the sake of buyer's time.

3.6 Sporas and Histos

Zacharia, Moukas and Maes [ZMM99] proposed Sporas as a reputational system implemented in an artificial electronic auction environment named Kasbah [GMCD97]. Kasbash is designed to provide a semi-automatic means of conducting business, with human users controlling a set of input variables for their agents. Sporas was inspired by the

foundations of the chess players' evaluation system called ELOS, which is a method to evaluate a player's relative strength in one-to-one games such as chess. Consequently it is not surprising that Sporas is an eBay-like-mechanism that, based on mutual evaluations after the transactions, provides global reputation values as a part of the identity of the agents and tries to ensure that the agents themselves cannot change these values. Histos in contrast uses the idea of social networking in order to calculate personalized reputation information and thereby makes use of the web of trust described earlier [Kuh99, p. 368 et seqq.]. Both ideas come about not as agent-specific (decentralized), but centralized from the logical storage point of view. Hence if an agent asks the systems for the reputation of another user, the system calculates this value (depending on the degree by which the user is involved in the social network) and reports it back to the inquiring agent. But how do Sporas and Histos work in detail?

Sporas is designed to make available reputation values to users of agent-based, loosely connected online communities [ZMM99, p. 3]. Thus the basic principle is a mathematical rating mechanism, with reputation being a natural number between 0 and 3000, whereas new agents start with a reputation value of 0, under which no agent can ever drop (as otherwise an incentive to simply change identities of the value falls under 0 would be given). Furthermore, if agents release reciprocal evaluation; then for each pair of agents only the very last rating is counted. The reputation of a single agent is then aggregated by the central system and published for all to see. As an agent's reputation gets higher, it is adjusted in a way that decreases the rate of accretion so that the rapidity of possible change (in both directions) decreases as reputation increases. The function used in order to calculate the new global reputation values for the agent after each transaction is as follows [ZMM99]:

$$R_{t+1} = \frac{1}{\theta} \sum_{1}^{t} \Phi(R_i) R_{i+1}^{other} (W_{i+1} - \frac{R_t}{D}) \tag{5}$$

with

$$\Phi(R) = 1 - \frac{1}{1 + e^{\frac{-(R-D)}{\sigma}}} \tag{6}$$

Thus, the new reputation value (R_{t+1}) is consistent with the sum of all previous reputation values, (recorded on a single context basis) weighted with the memory factor θ (the bigger θ , the bigger the memory of the system), modified by the attenuation function Φ (σ controls the slope of the function) set against the reputation of the evaluating agent (R_{i+1}^{other}) and the actual evaluation of the transaction (W_{i+1}) which comes from the inter-

val $[0.1, 1]$. The value D thereby represents the maximum value of reputation (3000). In case W_{i+1} is smaller than the hitherto existing reputation value divided by the maximum reputation D (this value should represent the expected value), the agent's reputation decreases; otherwise it increases. The goal of this formalism is that with a large number of evaluations, the reputation value R_{t+1} converges with the real reputation of the agent.[15] [MZM99, p. 316]

When trying to evaluate the mechanism, it should be said that some advantages are already present in this very standard system: (1) as agents' earlier evaluations are discarded when a new one is added; the illegal reciprocal reputation inflation between collusive agents can be antagonized. (2) Reputation can not descend below the entry level. This is a disincentive to get a new identity each time agents get to have a low reputation and hence could resolve the problem of "cheap pseudonyms". Furthermore it also acts as a discrimination imposed on new entrants. However, on the other hand, it discriminates against new-entrants as they are mostly going to have a lower level of reputation than defrauding agents, whose reputation value cannot decrease below the starting value of new-entrants (0). Another advantage of the Sporas system can be seen in the fact, that the reputation value can not exceed 3000. This is useful in avoiding accumulating permanent indestructible positive reputations. (3) The rater's reputation influences the degree to which its ratings are weighted. While the first two properties appear as viable solutions to some online reputation mechanism problems, the third makes good reputation more difficult to change than the initial low reputation. This is a preference for high reputation which is not encountered in every day, offline life, and which seems therefore unrealistic. In addition, reciprocal evaluation should not be considered the only way to implement a reputational mechanism, as this feature has a relevant impact on the type and quality of information thus produced.

Histos is Sporas' evolution and consequently has similar characteristics concerning the 5-stage-process model. It is a mathematical single-context model with a decentralized logical data storage and uses third party information in the recall phase as Sporas does. However, looking at the degrees of cheating accounted for, it has to be noted that - besides

[15] As the Sporas algorithm is supposed to be combined with Histos, it can be assumed that as in Histos, the evaluation of the underlying data structure is a graph in which agents are represented by nodes and the most current evaluations (including timestamps) are diagrammed by vectors (comparable to Schillo's TrustNet concept). Hence, in order to calculate the reputation value all vectors pointing to an agent have to be included chronologically. Thus, if a system consists of n agents, which in the worst case have all been evaluated by one another, the storage complexity of this structure is $O(n^2)$ and consequently very high.

the data storage - further differences between the two can be found.

Histos is used in an environment where the latter (or any other) system was already used to produce a bulk of evaluations. In this case it does not use the Sporas algorithm any more but uses the agents' social network analysis to weight the social evaluations received about possible partners. The reason for this expansion of the Sporas concept is that in Histos it is assumed that agents tend to trust evaluations of "friends" (agents they have already communicated with) more than the ones of complete strangers. [ZMM99, p. 166] That is why it is based on a collaborative filtering algorithm which counts previous exchanges' outcomes to create connection networks among "friends" and "friends of friends", up to any attainable level of connection. Thereby, as in Sporas, the net of paired evaluations is assumed to be a net of vectored graphs, consisting of nodes representing agents and vectors representing the most current evaluations by agent. In order to calculate the personalized reputation of an agent L, Histos examines the graph as to whether a direct path exists between A and A_L [ZMM99, p. 167]. In case a graph of length 1 can be found, A_L was evaluated by A. Otherwise a "breadth first" (i.e. a universal search parameter) search is conducted to find all paths between A and A_L, whereas the paths are maximally allowed to have the length N. For the calculation itself only the θ newest paths concerning the evaluation of A_L are being used. In order to finally calculate the reputation of A_L the reputation values of the agent one node part of have to be known. Therefore the reputation of these agents is calculated recursively (maximum length of the paths is consequently $N-1$) up to the agent one step apart from A, for which the evaluation by A can be used as a basis for the calculation. As a result the degree of complexity of the recursion is $O(\theta N)$. The actual calculation of the reputation value finally can be carried out with the following slightly modified formula[16] [ZMM99, p. 167]:

$$R_{t+1} = \frac{1}{\theta^t} \sum_{t-\theta^r}^{t} \Phi(R_{i+1})(R_{i+1}^{other} W_{i+1}) / \sum_{t-\theta^r}^{t} R_{i+1} \tag{7}$$

with

$$\theta^t = \min(\theta; m) \tag{8}$$

and

$$m = \deg(A_L) \tag{9}$$

[16] Regarding the application of the formula it has to be pointed out that it is only used if a direct path between two agents can be found. In case no or only paths longer than N between the two agents exist, the Sporas algorithm must be resorted to.

whereas m is the number of paths between A and A_L as discussed above. This alteration of the formula brings advantages in terms of the network load if agents had contacts with several other agents; however, the use of reciprocal evaluation which gave the input evaluations may have biases concerning the quality of social information expressed. Moreover, every evaluation is public and thus visible by the target; this may inhibit transmission of evaluations, as perceived responsibility increases. In any event, the weighing of evaluations for personal assessment of a target's reputation is important progress on the way of implementing an artefact's realism in online reputation mechanisms. Therefore, putting it in a nutshell, it can be concluded that Sporas / Histos can contribute to steady artificial communities by providing reputation mechanisms. However, a limitation of the mechanisms can be derived from the assertions that not the agents themselves, but their users, conduct the evaluations (what leads to subjectivity) and that the system only remembers the last mutual evaluations. This might lead to a collapse in the following situation: Assuming an agent B has a transaction with every other agent in the system and does not defraud, all other agents give him a positive evaluation. In a second round B again tries to barter with all other agents; however, this time it cheats every time and is consequently given negative evaluations. Now assuming that $\theta < 2N$ the following problem arises. As from every agent a path to $B < N$ is given, the question about B's reputation is answered with the help of the Histos algorithm. In this algorithm, however, the knowledge of B's negative evaluations does not help the single agent, as in the second round only the personal positive experience of each agent (from the first round) is considered. Consequently, in this very special case, the algorithm fails as due to their positive experience in the first round, every agent is going to trade with B in the second round, although in this round B is defrauding every time. The problem might be solved if not only the last 'personal' paired experiences of the agents were taken into consideration; however, this attempt reveals other problems of the centrally organized reputation mechanism as the calculation complexity as well as the storage complexity would increase significantly in case the last θ evaluations of and for every agent were stored. Hence it seems sensible to combine the Sporas / Histos approach with a "memory" of the agents as well as gossip-mechanisms and to leave the evaluation to the agent instead of their owners.

3.7 Yu and Singh

In the year 2000 Yu and Singh proposed a single-context agent-oriented model for social reputation/trust management which focused especially on electronic societies and MAS and uses the Dempster-Shafer theory of evidence[17] as the underlying computational framework [YS00]. In their papers Yu and Singh, on the one hand, introduced a gossip-mechanism ("If agent A encounters a bad partner B during some exchange, A will penalize B by decreasing its rating of B by β and informing its neighbours.") [YS00, p. 6] in which the gossip shall be transferred incrementally through the network of agents. On the other hand, it arranges for a mechanism that should help to include other agents' testimonies (witness information) in their own reputation calculations. Thus, in the trust part of the mechanism, Yu and Singh rely on the personal direct experience of agents concerning other agents. Thus, their agents store information about the outcome of every transaction they had with another agent and recall this information in case they are planning to bargain with an agent a second time. In case the agent meets another agent it has not traded with before and consequently does not have any direct information about this agent, the second part of Yu and Singh's model comes into play: the reputation mechanisms (based on third party information)[18]. In this mechanisms so-called referral chains are generated that can make witness information available across several intermediate stations. An agent is thus able to gain reputation information with the help of other agents in the network. However, this reputation information is not global as for example in eBay (where every user can see all profiles of all other members and every evaluation a user is given accounts for his reputation value), but depends on the referral chain the requesting agent is using. As this chain represent only a small extract of the whole network, the information delivered by the chains can be partial and thus may not be representative (decentralized logical information storage).

Yu's and Singh's algorithm itself finally works in two steps. The first one concentrates on updating the information an agent has about another agent after it has traded with it. It therefore classifies each transaction as a positive - (cooperation) or negative experience (defection) and respectively expresses a positive or negative evaluation for the other agent.

[17] The Dempster-Shafer theory is a mathematical theory of evidence based on belief functions and plausible reasoning which is used to combine separate pieces of information (evidence) to calculate the probability of an event. The theory was developed by Arthur P. Dempster and Glenn Shafer. More information including an in-depth explication of the theory can be found in [Sha90]

[18] It is important that the second part of the algorithm (the witness information based on the referral chains) is only used in the case that the agent does not have any direct information about his potential trading partner. Otherwise, it will only act based on its previous experiences.

In the case of a negative experience the evaluation is indicated in form of a value $\beta < 0$; in the case of a positive one in the form of a value $\alpha > 0$, whereas the scaling of the positive and the negative evaluations can vary and thereby directly influences the operation mode of the algorithm. In their paper Yu and Singh propose to choose (for example $\alpha = 0.07$ and $\beta = -0.5$) in order to factor negative experiences more strongly than positive ones. In order to finally update the reputation information after positive or negative interactions, Yu and Singh use the following mathematical formulas[19]:

Calculation of trustworthiness:

$$T_i(j)^t > 0; \quad j \quad has \quad cooperated \rightarrow T_i(j)^{t+1} = T_i(j)^t + \alpha * (1 - T_i(j)^t) \tag{10}$$

$$T_i(j)^t < 0; \quad j \quad has \quad cooperated \rightarrow T_i(j)^{t+1} = T_i(j)^t + \alpha * (1 - min(|T_i(j)^t|, |\alpha|) \tag{11}$$

$$T_i(j)^t = 0; \quad j \quad has \quad cooperated \rightarrow T_i(j)^{t+1} = T_i(j)^t + \alpha \tag{12}$$

$$T_i(j)^t > 0; \quad j \quad has \quad defected \rightarrow T_i(j)^{t+1} = (T_i(j)^t - \beta)/(1 - min(|T_i(j)^t|, |\beta|)) \tag{13}$$

$$T_i(j)^t < 0; \quad j \quad has \quad defected \rightarrow T_i(j)^{t+1} = T_i(j)^t + \beta * (1 - T_i(j)^t) \tag{14}$$

$$T_i(j)^t = 0; \quad j \quad has \quad defected \rightarrow T_i(j)^{t+1} = T_i(j)^t + \beta \tag{15}$$

In case no direct experience of the agent is available, it has to rely on the information given by the referral chain. This inclusion of witness information about an unknown agent n uses the following 3 variables:

L = number of different witness testimonies w in E=$\{E_{1w}, ..., E_{Lw}\}$[20]

[19] $T_i(j)^t$ is the trust of i in j at the time t.

[20] The calculation of the witness testimony is done with the help of the referral chains. In case of branching in the chain, the agent chooses the branch he assumes to be more trustworthy and then specifies the reputation values $T_i(j)$ (the trust of the witness in the target, the trust of the next-to-last in the witness, etc.) with the help of the operator $\otimes$ that is defined as follows:

, V = the subset of E that only includes the witness testimony of trustworthy witnesses and in case of equally trustworthy witnesses only includes the better testimonies, $\widehat{E}$ = mean value of all witness testimonies in V.

The calculation based on these variables works as follows:

$$T_i(n)^t > 0 \wedge \hat{E} > 0 \rightarrow T_i(n)^{t+1} = T_i(n)^t + \hat{E} * (1 - T_i(n)^t) \tag{16}$$

$$T_i(n)^t < 0 \quad XOR \quad \hat{E} < 0 \rightarrow T_i(n)^{t+1} = T_i(n)^t + \hat{E} / (1 - \min(|T_i(n)^t|, |\hat{E}|) \tag{17}$$

$$T_i(n)^t < 0 \quad and \quad \hat{E} < 0 \rightarrow T_i(n)^{t+1} = T_i(n)^t + \hat{E} * (1 + T_i(n)^t) \tag{18}$$

Inclusion of gossip ($T_k(n)$) about n that an agent i is told by k:

$$T_i(n)^t > 0 \quad and \quad T_i(k)^t > 0 \tag{19}$$

$$T_i(n)^{t+1} = T_i(n)^t + T_i(k)^t * T_k(n) * (1 - T_i(n)^t) \tag{20}$$

$$T_i(n)^t < 0 \quad and \quad T_i(k)^t < 0 \rightarrow T_i(n)^{t+1} = T_i(n)^t + T_i(k)^t * T_k(n) * (1 + T_i(n)^t) \tag{21}$$

$$\begin{gathered} different \quad algebraic \quad signs \rightarrow \\ T_i(n)^{t+1} = (T_i(n)^t + T_i(k)^t * T_k(n)) / (1 - \min(|T_i(n)^t|, |T_i(k)^t * T_k(n)|) \end{gathered} \tag{22}$$

In the context of several extensive experiments, Yu and Singh used these equations and showed that the implementation of their mechanism results in a stable system in which the reputation of defrauding agents decreases rapidly while the new, cooperating agents experienced a slow, but almost linear increase in reputation, if suitable values for α and β where chosen (i.e. $|\alpha| < |\beta|$). Although the experiments indicate the functionality of the algorithm, two problems have to be taken into consideration. The first one was already

---

$$x \otimes y : \begin{cases} xy & \text{for } x \geq 0 \wedge y \geq 0 \\ -xy & \text{else.} \end{cases}$$

mentioned above: the model does not combine the sources of information it takes into account (direct information and witness information [including gossip]) so that if an agent has already traded with another one it cannot use the network information any more and thus might need an unnecessarily long time to react to an agent defecting suddenly who cooperated before. Furthermore Singh and Yu do not give any explanations of how their agent-centered storage of information of the social (i.e. public) knowledge (for example of the referral chains) is supposed to be organized. Consequently no analysis of the network-load and the storage-intensity can be made.

3.8 Padovan et al. (AVANLANCHE)

The Avalanche reputation mechanism [Eym00] that was implemented in Java is a single context model that uses a mathematical rating mechanism. It includes domain-specific "rating agents" (or -agencies) which can each provide specific reputation information [PSEP02]. In Avalanche the single agents pursue the goal of trading goods in several interlinked market-places. As a part of their negotiation strategy they therefore use an agent-specific reputation coefficient between 0 (bad reputation) and 1 (good reputation) which they assign to every other agent they know. This coefficient is then used by the agents to compare possible transaction partners not only based on prices, but on the possible loss as well. Therefore the reputation value the agent attributes to his possible trading partner is accounted for as the probability of the non-advocacy of a loss, so that in case of a 100 per cent assumption of a loss, twice the price is used as a bargaining basis [Pad00, p. 9 et seqq.]. After each transaction the participating agents then change their mutual evaluations. Thus, a successful transaction is rated with $r = 1$ whereas a failed one is rated with the value $r = 0$. For the final calculation of the reputation coefficient of agent Y (from agent X's point of view) Avalanche uses the following formula which furthermore includes the weighting factor α:

$$R^X_{Y_{t+1}} = R^X_{Y_t}(1 - \alpha) + r_t\alpha \tag{23}$$

One important feature of this equation is that cooperative and non-cooperative behaviour are both represented by α (or -α), although cognitive conceptualizations of reputation and trust suggest to punish defection heavier (cp. Yu and Singh for example). Concerning the reputation calculation of unknown agents, Padovan et al. propose to take the average value of all known reputation evaluations of all agents, or, in case these data are not available,

to use a default value. The introduction of one or more rating-agencies is discussed as an expansion of the model as it stands so far. These agencies are a kind of "subjective" external elements to which single agents have to (or can) forward their evaluations. For every agent Y the agency then saves a reputation value R_Y that is changed analogue to the changes of the reputation coefficients by the agent. However, when it comes to collecting the decentralised stored reputation information, the agency does not save the originator of the single evaluations, but only stores the actual reputation value and its variance. Hence, an agent can constantly deliver evaluations for another agent and thus can directly influence his reputation. Thus in theory two agents can push their mutual coefficients in the agencies data base. For the calculation of changes of the agent-specific central reputation value $R_{Y_{t+1}}$ after an evaluation by X (r_t^X) the following formula is used, whereas γ represents the agent-specific equivalent to α in the first formula.

$$R_{Y_{t+1}} = R_{Y_t}(1-\beta) + r_t^X \beta; \quad with \quad \beta = \gamma R_{X_t} \text{ [21]} \tag{24}$$

In summary, an agent in Avalanche, on the one hand, has the possibility to consult his own experiences for the evaluation of potential trading partners, and, on the other, can resort to the combined reputation values by the central rating agencies, which however are not fraud resistant, as they do not consider the possibility of incorrect information (level 1 in the modification phase). In addition, Avalanche does not have any direct social component (for example comparable to a TrustNet).

Comparing Avalanche to the concepts of Marsh and Yu and Singh that proposed the co-domain $[-1, 1]$, it attracts attention that it used $[0, 1]$ as Schillo does, instead and thus emanates from the theory of probabilities. Several points speak in favour of the approach - above all the relatively simple mathematical background. However, from the cognitive point of view it remains questionable whether trustworthiness can be quantified so simple (and possible linear), as between trust and distrust as well as between the different trust levels huge differences exist [Luh89]. Furthermore Padovan et al. make no comments about the case in which an agent has a reputation level below 0.5 and consequently defects more often than cooperates. Under the assumption that the price is cleared with the risk of loss, a defecting agent could participate in transactions a good many times (depending on the choice of α, as well as on the factor whether only personal or agency information is chosen). It only needs to offer a price that is low enough and as it has no intentions

[21] In case information about software agent Y are known in the agency, the mean value of all known reputation values is used instead of R_{Y_r}.

of providing the good or service it does not have any costs. With regard to the rating agencies another problem arises. Thus in the papers of Padovan et al. [Pad00, PSEP02] it remains unclear how it should be implemented in the Avalanche structure. Should there be an agency for every market or is a system-wide agency intended? Or might have Padovan though of several competitive agencies and, if this is the case, how do the agents choose the agency they rely on? In addition it remains unclear whether agents have to pay for the information from the agency or whether they might get any money for the evaluation they contribute. Hence, although the idea of a combination of own evaluations and a rating agency seems relatively flexible and easy to realise, the weaknesses of the concept have to be thought about and cognitive approaches such as gossip mechanisms should be implemented to make the model more realistic.

3.9 Foner

Another reputation mechanisms is the one Leonard Newton Foner proposed in his PhD thesis. Foner's approach is a single context mechanism that shares similarities with PGP and the "Web of Trust" [Fon99, p. 43], [Kuh99, p. 358 et seqq.]. Foner introduces the matchmaking system *Yenta* and in this context discusses the question of trust with special focus on *TRUSTe*[22] with respect to privacy adherence. According to his papers, it is Foner's goal to develop an agent-based system which allows the pooling of users with similar fields of interest without endangering privacy-relevant information unnecessarily. Therefore Foner describes a decentralized system architecture that allows for the cooperation of several agents who are not necessarily familiar with one another, while maintaining high security standards. *Yenta* is one sample application for this idea. It relies on recommendations articulated by local interest-clusters (collaborative filtering). To give any estimations about the trustworthiness of the users *Yenta* furthermore contains a reputation mechanism, which despite the decentralised system architecture, is centralized when it comes to the storage of the cooperative behaviour. This mechanism consists of self-portrayals of users which are known to the local *Yenta* program, and whose trustworthiness is guaranteed by third parties with encrypted digital signatures. In order to finally find out the trustworthy agents, *Yenta* uses uses third party information in form of a kind of web of trust in the recall stage. Hence, first of all, an agent which tries to estimate the

[22] *TRUSTe* was founded in 1997 by the Electronic Frontier Foundation (EFF) and Commerce.net to foster online commerce by helping businesses and other online organizations to self-regulate privacy concerns. Today it runs the world's largest privacy seal program, with more than 2,000 Web sites certified, including IBM, the Oracle Corporation or eBay.

trustworthiness of an unknown agent checks whether it knows the trustworthiness of the agent signing the self-portrait. If this is not the case, the search is expanded one step, etc. Foner compares this to "small-town-gossip" [Fon99, p. 44] with the distinction that in *Yenta* self-portraits are the main information source, whereas in the latter third party information is more important. Thereby the final character of the network structure (only some local reputation spots or a real network) is dependent on the social and political decisions of the users, and is not bound to the system architecture [Fon99, p. 44]. In contrast to most of the other mechanisms and proposals reviewed so far, when it comes to the second stage of the 5-stage-process model, Foner does not see trustworthiness as an expected value of the future behaviour or a measure between -1 and +1, but as an explicit picture of social networks which have already influenced the Histos idea.

3.10 Sabater and Sierra (ReGreT)

Developed by Sabater and Sierra [Sab03, SS01, SS02], *ReGreT* is a reputation mechanism which studies reputation in complex societies. It is an interesting model as it accounts for a multi-faceted concept of reputation. In fact, it identifies three levels for which reputational variables should be considered: individual, social and ontological system-related factors. The individual one is considered the most reliable, as it stems from (reciprocal) evaluation resulting from direct interaction with the target. [KKW07] When the agent does not have enough direct information, it should use the social facet of reputation, which is made up of three components: "witness reputation"; "neighbourhood reputation"; and "system reputation":

Witness information refers to evaluation coming from direct experience, and then referred by a fellow agent. In order to choose whom to ask for information, the past's partners' evaluations about the target are considered using a heuristic that groups correlated evaluations (coming from the same interactions) and filters them. Then one agent is chosen and its evaluation of the target is picked up. This kind of information is not as trustworthy as the direct information mentioned above. Every piece of witness information has some uncertainty around itself. More detailed and extensive approaches, like the advancement of ReGreT, called Repage [SPC06], have different modules to specify these two sources of witness reputation. ReGreT does not differentiate between these two sources.

Neighbourhood reputation is the result of social contagion. In fact, the system adjusts the previous reputational value to account for the reputation of the group that surrounds the

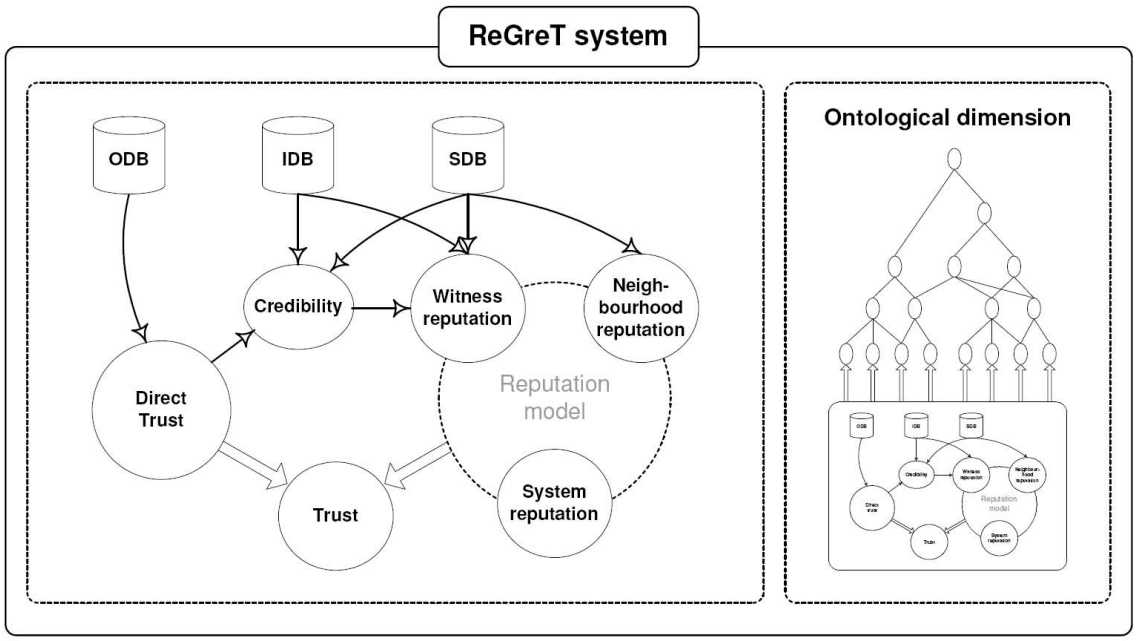


Figure 5: The ReGreT Approach [Sab03, p. 42]

target, which is calculated from their relational behaviour.

Finally, *system reputation* is derived from the "institutional role" assigned to the target by the system. If more than one role is assigned, only the prominent one is considered. The roles may be those of seller, buyer, or the affiliation to one known company, of which the subject has a reputation as it is, perhaps, affiliated to a concurrent company. The ontological specification of reputation is used to distinguish different features of the target's possible behaviour such as delivery time, price surcharge, quality swindling, and so on. Each of them is assigned a value that describes the target in detail. So, different domains of reputation might thus be accounted for. The order of listing of the components reflects their descending weight to the final value of the target's reputation. This value is a probabilistic one, describing the likelihood that the target will cooperate along the logical dimensions of the contract specified by the ontology of it.

Following the three information sources, the ReGreT model employs three data bases, the "outcomes data base" (ODB) for own experiences with other agents, the "information data base" (IDB) for storage of witness information, and the "sociograms data base" (SDB) to represent social structures (see figure 5).

Regret introduces an important distinction among different sources of social evaluation. Its basic distinction between direct (personal) evaluation and the socially transmitted one reflects the distinction between image and reputation; a fundamental one in the social

cognitive view of the artefact. In addition it places agents in a social environment, which is useful to account for social contagion as a necessary part of reputation. However, the authors recognize that the sociological subdivision used in the model is probably incomplete; it may in any event result in a suitable compromise between model complexity and the mechanism's requisites. Nevertheless it has to be stressed that social contagion is just a part of the phenomenon that should not be overly-considered. Moreover, the model seems to give way to communication between agents only as a result of asking, while in real life it is as well the case that social information is received as a spontaneous offering, a sort of "gift". The important point not considered is the moment of spreading social information, also called memetic act. Should it always be reciprocal evaluation? This may have impact on the quality of information produced, as set of the evaluators overlaps set the set of the target agents (i.e. the ones being evaluated); this may give way to benevolence in information dissemination toward the target, resulting in the application of a "rule of courtesy" and therefore the production of an over-inflated reputation (or low provision of it). In this sense, the model is not realistic. The realism of modelling is important to achieve the same benefits in place in the human artefact, which seems to function rather well with regard to the task of assuring distributed social control and norm enforcing.

3.11 Sabater et al. (RepAge)

RepAge [SPC06] is a decentralized single context computational reputation system, that does not purely rely on mathematical ratings alone, but models the cognitive processing of reputational information in the mind of an agent, using the cognitive theory of reputation covered by [eRe06, CP02, MC00], as well as ideas from the Regret systems and adds some new elements to this. Thereby Sabater et al. especially focus on the difference between image and reputation, which in their view suggests "a way out from the paradox of sociality, i.e. the trade-off between agents' autonomy and their need to adapt to social environment" [SPC06, abstract]. This is explained by the authors by highlighting the importance of two aspects with regard to the concept of reputation: on one hand, agents are autonomous if they select partners based on their social evaluations (*images*), and on the other hand, they need to update evaluations by taking into account others'. Hence, Sabater et al. draw the conclusion that social evaluations must circulate and be represented as "reported evaluations" or "meta-evaluations" (i.e. *reputation*), in order for

agents to decide whether to accept them or not and whether to integrate this information with their own image of the target. Consequently, RepAge considers possible defrauding behavior of information sources in the recall phase.

RepAge itself is based on an algorithm which endows the agents with a heuristic used for the processing and integration of the different components of reputation within a single mind when evaluating and rating other, thus pursuing a cognitive approach in this respect. It considers social evaluation as fuzzy values of 5 levels, and thereby tries to model actual people's informal evaluative statements by weighting the aggregation of this fuzzy values. The aggregation itself is done with the help of a formula proposed by [Jag04] to which the authors added the calculation of strength in order to get rid of the problem the formula looses it sense in case the denominator turns 0 (what can happen easily):

$$w_i = \frac{\prod_i w_i^j}{\sum_i \prod_j w_i^j} \quad [23] \tag{25}$$

Based upon these considerations, the authors present the an architecture for their reputation mechanisms that can be seen in figure 6:

As it can be seen in the figure, RepAge very much focuses on the individual agents and thus all information storage as well as all reasoning is done decentralised in the individuals' memories. The memory itself works in a bottom-up fashing, i.e. resulting from the communicated images and meta-evaluations (reputation information) that are attributed to the contractual context (bottom layer) decision are being made. Thereby it is important to notice that at this point of time the predicates (i.e. information derived from different sources) are not valued by RepAge at this stage. Based on the the information transactions are being made and their outcome accounted for at the next layer. This is not only done on a binary basis (i.e. the outcome is not just a tuple contract-fulfilment with fulfilment taking the value 0 or 1) but an evaluation is done on how the contract was fulfilled (i.e. the quality of service (QoS) is being accounted for).

Last but not least, in this conceptual layer, one final step needs to be taken in the RepAge model: the evaluation on the information sources. This is done by comparing the output of the transaction with the information given by the sources beforehand. Based on the results of the evaluation, the model, in the next stage derives at five types of predicates:

[23] In the formula w_i^j represents the weight of the fuzzy evaluations whereas the lower index i refers to the different weights of the same fuzzy evaluations, and the higher one j is used to distinguish the evaluations to aggregate.

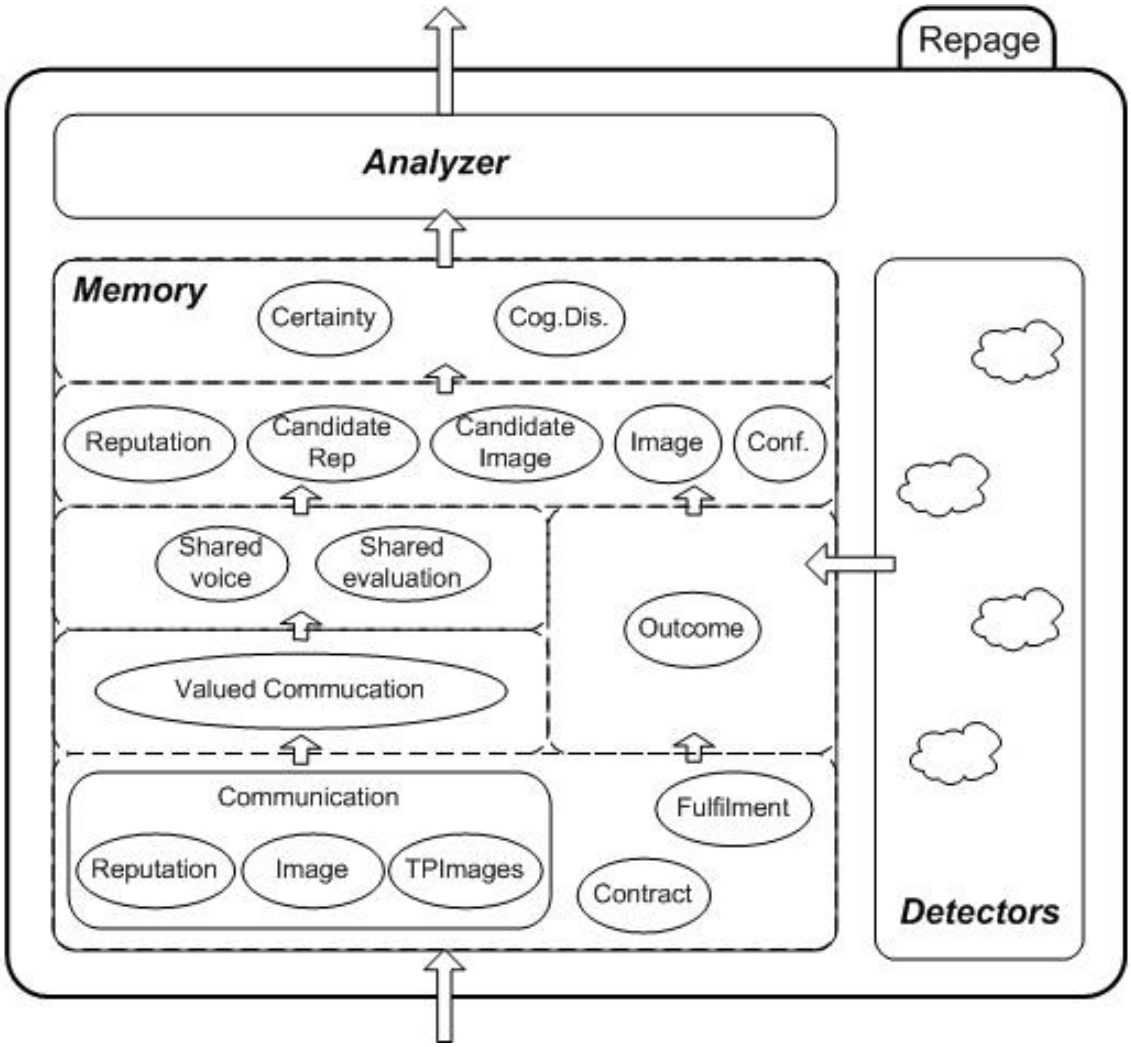


Figure 6: The RepAge Architecture [SPC06]

Candidate Image, *Candidate Reputation*, *Image*, *Reputation* and *Confirmation*. As the prefix "candidate" already indicates, of the five the first two have no sufficient support to become real image or reputation information yet. This might for example be the case if the elements contributing are not of a sufficient number. The last predicate "confirmation" comes from the former layer. It accounts for the quality of the information the agent received by others. Thus it is a value of how accurate the information provided by the informers was (based on the results of the transaction).

Finally, the last layer consists of two predicates: *Cognitive dissonance* and *Certainty*. Whereas "certainty" refers to a state where the individual is certain about pieces of information that are all in the same line of reasoning, "cognitive dissonance" refers to a situation where the pieces of information all have sufficient support, however they contradict with regard to a specific target. Depending on how strong and relevant this dissonance is, the

model accounts for actions to be taken to solve the dissonance.

Once all information from one transaction have passed through all the stages and layers of the agents memory, the process of starts over and over again for each new piece of information. Thus, the memory (including the top layer) are constantly updated giving the agent the ability to learn and reason about its behaviour as well as its own strategy.

4 Summary

In this paper, we have proposed a classification scheme for analysing reputation mechanisms and comparing them. This scheme consists of five stages that form a process and link two transactions. In order to show how different models are to be seen in the context of our 5-Stage-process model and which design choices could be made within the five stages, in the further course of this paper we then have analysed a set of well known reputation mechanisms. Thereby it has to be noted that the choice of mechanisms was made to show the differences in design, and consequently the list of reputation mechanisms we examined does not lodge the claim of completeness.

To sum up, as a last step, the results of the analysis shall now be presented in a condensed form using a summary table (see table 4). Of course, the heterogeneity as well as complexity of the models presented makes the condensed comparison very difficult, especially as only a small number of criteria can be accounted for. Consequently a slight subjective input that all kinds of categorizations like this one have, remains. Furthermore, it must be said, that if not stated differently, we only concentrated on the initial models, of which many however have been developed further by the authors themselves or by others.

Nevertheless now we will try to condense and compare the mechanisms discussed. The criteria for the comparison are all based on the four primary stages of the process-model and should therefore need no further explanation. The can be seen in the following table:

Recording of cooperation behaviour	single-context model (**SC**) multi-context model (**MC**)
Rating of cooperation behaviour	cognitive (**C**) mathematical (**MA**)
Storage of cooperation behaviour	centralized storage (**CS**) decentralized storage (**DS**)
Recall of cooperation behaviour	trust model (**T**) reputation model (**RE**)
Degree of cheating accounted for in the recall phase	Level 0 (**L0**) Level 1 (**L1**) Level 2 (**L2**)

Table 3: Summary of the 5-Stage-Process-Model Criteria

Using these criteria, for the mechanisms discussed in this paper, the following final picture can be drawn:

	Recording	**Rating**	**Storage**	**Recall**
Marsh	MC	MA	DS	T, no information
Schillo	SC	MA	DS	RE, L2
Rasmusson and Janson	SC	MA	CS	RE, assumes cheating, but paid agents are assumed to tell the truth
Abduhl-Rahman and Hailes	MC	MA	DS	RE, L0
Regan and Cohen	MC	MA	DS (with CS components, e.g. "advisors")	RE, L2
Sporas	SC	MA	CS	RE, L0
Histos	SC	MA	DS	RE, L2
Yu and Singh	SC	MA	DS	RE, L2
AVALANCHE	SC	MA	DS (with CS comp., e.g. "rating agencies")	RE, L0
Foner	SC	MA	CS	RE, L0
ReGreT	MC	MA	DS	RE, L2
RepAge	SC	CO	DS	RE, L2

Table 4: Classification of the Models

As expected and described before, most of the models are single-context models that use several sources of information, with the most important one being witness information. Thereby the focal point in the information processing are mathematical paradigm, with only one mechanisms focusing on cognitive information. This is especially interesting as an correlation between the rating and the strategy adaption seems to exist. Thus, although most mathematical models account for the evolution of the starting trust and/or reputation parameters, a adaption of the actual strategy changing the formulas used, cannot be found in any of the mathematical models. Concerning the data storage, most mechanisms focus on centralized solution, however first attempt for decentralized approaches can be found. Last but not least, due to the focus of this paper, all except one mechanism are reputation mechanisms, accounting for the importance of third party information. However, about only half of the reputation mechanisms account for the problem that information sources may deliver wrong information on purpose making, which however makes these models far more complex.

Summing up, reputation and trust mechanisms with manifold facets of implementation can be found when reading current literature on securing artificial societies. These range from simple trust model to very complex cognitive approaches that take into account cheating and much more. All of the models have their particular advantages and disadvantages allowing the designers of artificial societies to choose a mechanisms based on their needs (e.g. in terms of complexity). Nevertheless, when choosing one needs to keep in mind that although being presented with the help of different stages in this paper, these stages are interlinked. Thus, design choices in one stage might influence another stage, so that reputation as process should always be reasoned about.

References

[Ake70] George Akerlof. The market for 'lemons'. quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84:488–500, 1970.

[ARH97a] Alfarez Abdul-Rahman and Stephen Hailes. A distributed trust model. In *NSPW '97: Proceedings of the 1997 workshop on New security paradigms*, pages 48–60, New York, NY, USA, 1997. ACM.

[ARH97b] Alfarez Abdul-Rahman and Stephen Hailes. Using recommendations for managing trust in distributed systems. In *IEEE Malaysia International Conference on Communication*, 1997.

[ARH00] Alfarez Abdul-Rahman and Stephen Hailes. Supporting trust in virtual communities. In *HICSS*, 2000.

[AS03] D. Ariely and I. Simonson. Buying, bidding, playing, or competing? value assessment and decision dynamics in online auctions. *Journal of Consumer Psychology*, 13(1):113–123, 2003.

[CdRF97] C. Castelfranchi, F. de Rosis, and R. Falcone. Social attitudes and personalities in agents. In *Proceedings of the AAAI Fall Symposium on Socially Intelligent Agents*, Cambridge, Massachusetts, 1997.

[CP02] Rosaria Conte and Mario Paolucci. *Reputation in Artificial Societies: Social Beliefs for Social Order*. Springer, October 2002.

[eRe06] eRep. Review of internet user-oriented reputation applications and application layer networks. http://megatron.iiia.csic.es/eRep/files/eRep_D1.1_ReviewInternetReputation.pdf, September 2006.

[eRe07] eRep. E-institutions oriented to the use of reputation. http://megatron.iiia.csic.es/eRep/files/eRep_D2.1_eInstitutionsReputation.pdf, May 2007.

[Eym00] Torsten Eymann. *AVALANCHE: Ein agentenbasierter dezentraler Koordinationsmechanismus für elektronische Märkte*. PhD thesis, Albert-Ludwigs-Universität, 2000.

[Fon99] Leonard Newton Foner. *Political artifacts and personal privacy: the yenta multiagent distributed matchmaking system.* PhD thesis, Massachusetts Institute of Technology, 1999.

[Gam90] Diego Gambetta. Can we trust trust? In Diego Gambetta, editor, *Trust: Making and Breaking Cooperative Relatioins*, pages 213–237. Published Online, 1990.

[GMCD97] R. H. Guttman, P. Maes, A. Chavez, and D. Dreilinger. Results from a multiagent electronic marketplace experiment. In *Second International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology*, 1997.

[Jag04] R. Jager. On the determination of strength of belief for decision support under uncertainty – part ii: fusing strenghts of beliefs. *Fuzzy Sets and Systems*, 142(1):117–128, 2004.

[JIB07] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.

[KKW07] Stefan König, Sven Kaffille, and Guido Wirtz. Implementing regret in a decentralized multi-agent environment. In Paolo Petta, Jörg P. Müller, Matthias Klusch, and Michael P. Georgeff, editors, *MATES*, volume 4687 of *Lecture Notes in Computer Science*, pages 194–205. Springer, 2007.

[Kuh99] Rainer Kuhlen. *Die Konsequenzen von Informationsassistenten. Was bedeutet informationelle Autonomie oder wie kann Vertrauen in elektronische Dienste in offenen Informationsmärkten gesichert werden?* Suhrkamp, Frankfurt am Main, 1999.

[Luh89] N. Luhmann. *Ein Mechanismus der Reduktion sozialer Komplexität.* Ferdinand Enke Verlag, 1989.

[Mar94] Stephen P. Marsh. *Formalising Trust as a Computational Concept.* PhD thesis, Department of Mathematics and Computer Science, University of Stirling, April 1994.

[MC00] M. Miceli and C. Castelfranchi. The role of evaluation in cognition and social interaction. In K. Dautenhahn, editor, *Human cognition and social agent technology*. Benjamins, Amsterdam, 2000.

[MGM06] Sergio Marti and Hector Garcia-Molina. Taxonomy of trust: categorizing p2p reputation systems. *Computer Networks*, 50(4):472–484, 2006.

[MZM99] A. Moukas, G. Zacharia, and P. Maes. Amalthaea and histos: Multiagent systems for www sites and reputation recommendations. In M. Klusch, editor, *Intelligent Information Agents. Agent-Based Information Discovery and Management on the Internet*. Springer, Berlin, Heidelberg, New York, 1999.

[Pad00] Boris Padovan. Ein vertrauens- und reputationsmodell für multi-agenten systeme, 2000.

[PEJ+09] Mario Paolucci, Torsten Eymann, Wander Jager, Jordi Sabater-Mir, Rosaria Conte, Samuele Marmo, Stefano Picascia, Walter Quattrociocchi, Tina Balke, Stefan König, Thijs Broekhuizen, Debra Trampe, Mirjam Tuk, Ismel Brito, Isaac Pinyol, and Daniel Villatoro. Social knowledge for e-governance: Theory and technology of reputation. Technical report, ISTC-CNR, Rome, Italy, 2009.

[PSEP02] B. Padovan, S. Sackmann, T. Eymann, and I. Pippow. A prototype for an agent-based secure electronic marketplace including reputation tracking mechanisms. *International Journal of Electronic Commerce*, 6(4):93–113, 2002.

[Ras96] Lars Rasmusson. Socially controlled global agent systems. Master's thesis, Royal Institute of Technology, October 1996.

[RC05] Kevin Regan and Robin Cohen. Indirect reputation assessment for adaptive buying agents in electronic markets. In *Proceedings of the Business Agents and Semantic Web (BASeWEB05)*, Victoria, Canada, 2005.

[Rip01] M. Ripeanu. Peer-to-peer architecture case study: Gnutella network. Technical report, University of Chicago, 2001.

[RJ96] Lars Rasmusson and Sverker Jansson. Simulated social control for secure internet commerce. In *NSPW '96: Proceedings of the 1996 workshop on New security paradigms*, pages 18–25, New York, NY, US, 1996. ACM.

[RRJ97] L. Rasmusson, A. Rasmusson, and S. Janson. Using agents to secure the internet marketplace - reactive security and social control. In *Practical Applications of Agents and Multi-Agent Systems 1997 (PAAM'97)*, 1997.

[Sab03] Jordi Sabater. *Trust and Reputation for agent societies.* PhD thesis, Institut d'Investigació en Intel.ligència Artificial (IIIA), 2003.

[Sch99] M. Schillo. Vertrauen und betrug in multiagentensystemen: Erweiterung des vertrauensmodells von castelfranchi und falcone um eine kommunikationskomponente. Thesis, Lehrstuhl Deduktion und Multi-Agenten Systeme, Universität des Saarlandes, 1999.

[SFR00] Michael Schillo, Petra Funk, and Michael Rovatsos. Using trust for detecting deceitful agents in artificial societies. *Applied Artificial Intelligence*, 14(8):825–848, 2000.

[Sha90] Glen Shafer. Perspectives on the theory and practice of belief functions. *International Journal of Approximate Reasoning*, 4(5-6):323–362, 1990.

[SPC06] J. Sabater, M. Paolucci, and R. Conte. Repage: Reputation and image among limited autonomous partners. *Journal of Artificial Societies and Social Simulation*, 2(2), 2006. http://jasss.soc.surrey.ac.uk/9/2/3.html.

[SS01] J. Sabater and C. Sierra. Regret: A reputation model for gregarious societies. In *Fourth Workshop on Deception Fraud and Trust in Agent Societies, Montreal, Canada*, pages 61–70, 2001.

[SS02] J. Sabater and C. Sierra. Social regret, a reputation model based on social relations. *SIGecom Exch.*, 3(1):44–56, 2002.

[SS05] J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artificial Intelligence Review*, 24(1):33–60, 2005.

[SW05] J. D. Sonnek and J. B. Weissman. A quantitative comparison of reputation systems in the grid. pages 242–249, 2005.

[Szt99] Piotr Sztompka. *Trust: a sociological theory.* Cambridge University Press, 1999.

[TC01] T. Tran and R Cohen. A learning strategy for economically-motivated agents in market environments. In *IJCAI01 workshop on knowledge discovery from distributed, dynamic, heterogeneous, autonomous sources*, pages 51–56, 2001.

[Win99] M. Winter. The role of trust and security mechanisms in an agent-based peer help system. In *Autonomous Agents '99, Workshop on Deception, Trust, and Fraud in Agent Societies*, pages 139–149, Seattle, Washington, United States, 1999.

[YS00] Bin Yu and Munindar P. Singh. A social mechanism of reputation management in electronic communities. In *CIA '00: Proceedings of the 4th International Workshop on Cooperative Information Agents IV, The Future of Information Agents in Cyberspace*, pages 154–165, London, UK, 2000. Springer-Verlag.

[Zac99] G. Zacharia. Trust management through reputation mechanisms. In C. Castelfranchi, R. Falcone, and B. S. Firozabadi, editors, *Deception, Fraud and Trust in Agent Societies*, pages 163–167. Seattle, 1999.

[ZMM99] G. Zacharia, A. Moukas, and P. Maes. Collaborative reputation mechanisms in electronic marketplaces. In *Proceedings of the 32nd Hawaii International Conference on System Sciences, Wailea Maui*, 1999.

The Internet has caused a revolution in trading. Especially cheap items are now easy to buy and sell on the Internet. As a consequence, sellers nowadays offer a wide range of products on the web, creating an abundance of choice for consumers. Consumers have the opportunity to browse on different auction sites for the item they really want. Along with this success story, however, came the stories of people being cheated by fraudulent online sellers. These frauds cover a range from not delivering what has been promised, the overrating of a product's condition, to deliberate acts of theft. They are a result of so-called asymmetric information. Trust and reputation mechanisms are intended to address this asymmetric information distribution. This article surveys the most common trust and reputation systems.

ISSN 1864-9300