Lehrstuhl für
Wirtschaftsinformatik
Information Systems
Management

No. 48

November 2009

Bayreuther Arbeitspapiere zur Wirtschaftsinformatik

Miriam Heitz and Stefan König

Reputation in Multi Agent Systems and the Incentives to Provide Feedback

Bayreuth Reports on Information Systems Management

UNIVERSITÄT
BAYREUTH

ISSN 1864-9300

Authors:

Miriam Heitz and Stefan König
University of Bayreuth
stefan.koenig@uni-bayreuth.de

Information Systems Management
Working Paper Series

Edited by:

Prof. Dr. Torsten Eymann

Managing Assistant and Contact:

Raimund Matros
Universität Bayreuth
Lehrstuhl für Wirtschaftsinformatik (BWL VII)
Prof. Dr. Torsten Eymann
Universitätsstrasse 30
95447 Bayreuth
Germany

Email: raimund.matros@uni-bayreuth.de

ISSN 1864-9300

Contents

List of Figures

List of Tables

Abbreviations

MAS	Multi Agent System
ORep	Overall Reputation
P2P	Peer-to-Peer
QoS	Quality of Service
QoE	Quality of Experience
R-Agent	Broker Agent
RRep	Recommendation Reputation
SRep	Service Reputation

1 Introduction

The emergence of the internet leads to a vast increase in the number of interactions between parties that are completely alien to each other. In general, such transactions are likely to be subject to fraud and cheating. If such systems use computerized rational agents to negotiate and execute transactions, mechanisms that lead to favorable outcomes for all parties instead of giving rise to defective behavior are necessary to make the system work.

Think of e-commerce systems in which completely rational agents automatically search for providers and negotiate terms of trade after detecting a need. Since these agents are set-up to maximize the profit of the party they are acting for, they will cheat on their trading partners and refrain from paying for services that have already been delivered, if the rules of the game are not designed in such a way that cheating reduces their expected future gains from trade.

To the extent that the framework the actors operate in is badly designed, it is likely to resemble a prisoners dilemma. Once-off interactions in prisoner's dilemmas lead to mutual defection and therefore destroy the very grounds that make trade worthwhile. Reputation mechanisms can play a major role in making reliable promises between rational and anonymous actors possible. Such systems transform once-off interactions between agents in iterated interactions, and hence make cooperation a rational strategy.

Reputation mechanisms need feedback from the agents engaged in trade. Unfortunately, it is not in the best interest of a rational agent to report feedback, since that would provide a competitive advantage to the other agents. Suppose, for example, that a trading partner cheated on an agent. Why should this agent report the cheating? If he competes with the agents that would benefit from the report, he would provide them with valuable information that gives them a competitive advantage. If, on the other hand, the interaction went well, and the agent gave positive feedback, that would increase the reputation of the trading partner and therefore diminish its own. In an unregulated environment, therefore, it is not rational for agents to report feedback either way.

In consequence, a trust establishing mechanism has to be implemented with two features: first, it has to encourage rational agents to give honest feedback. Second, it has to serve as a tool to communicate hidden characteristics of and feedback about the transaction partner's behavior. This paper will discuss several trust and reputation mechanisms that show these characteristics. This paper does not intend to give a complete exhausting survey of trust and reputation models. Therefore we refer for example to [AG07, BKE09, JIB07, MGM06, RKZF00, SS05]

This paper is organized as follows: Chapter 2 on the facing page presents the necessary background and related work on reputation mechanisms in off-line and on-line settings. Chapter 3 on page 7 will exemplify five incentive setting reputation mechanisms for multi agent environments. Finally, chapter 4 on page 31 concludes with an overview about the reputation mechanism and the incentives that are necessary to make it rational to report feedback truthfully.

2 Reputation

The trust problem described in the introduction makes it necessary to design reputation mechanisms for multi agent market places [KHPE08, p. 1]. Service providers and consumers take the risk of a defecting partner; reputation mechanisms promise to signal whether a partner is trustworthy or not. They can facilitate "to promote cooperative and honest behavior among self-interested economic agents" [Del05, p. 210].

Their objective is to provide information about hidden characteristics e.g. quality of service for other community members. The mechanism has both a sanctioning and a signaling function. It signals if a service provider has delivered in past transactions and sanctions bad behavior of consumers and providers such as not paying or not delivering respectively with bad reputation values. Future partners can estimate how likely it is that a transaction will be successful with the help of the reputation value. Derived trust through feedback from other market participants can decrease the risk of the lack of trust. Thus reputation is needed as an indicator for the trustworthiness of the partner and the quality of the service.

Another problem in multi agent societies is that rational agents will not report feedback at all when it is not rational for them to do so. [Del05] names two reasons. First, published feedback is a public good which can be used by everyone at no cost. Hence, the agent giving feedback will not benefit from his task. Since any agent giving feedback is experiencing costs but no gain, no one has an incentive to give feedback. Secondly, in order to provide feedback one of the agents has to take the risk of interacting with another agent without having information about his past behavior [Del05]. Thus the implemented reputation mechanism needs to set incentives for users to submit feedback and additionally has to further trustworthy feedback. In chapter 3 on page 7 such incentive setting reputation mechanisms which make it rational for agents to report and do so truthfully will be introduced.

Section 2.1 on the following page will first present the process of reputation formation in general. Section 2.2 on page 5 will look at reputation in on-line environments specifically. The preconditions needed for reliable and trustworthy feedback will be presented in section 2.3 on page 6.

2.1 Reputation Formation and Word-of-Mouth

Reputation has been a powerful tool for a long time. Büschken looks at reputation networks in off-line settings and develops a model of reputation building. First of all word to mouth recommendations exist whenever a consumer tells another person about his experience. This message will be called image [CP02, p. 72]. Image is important since it is the basis for reputation. After the formulation of an image it diffuses in the market and is thereby objectified. The following illustration shows the process.

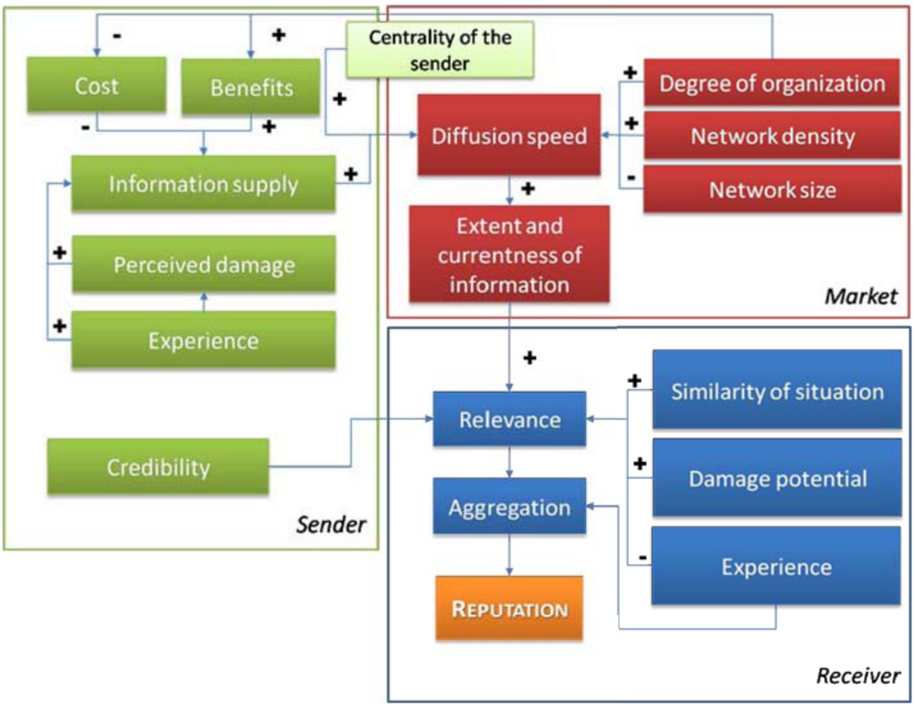


Figure 1: Reputation Formation (following [Bü00])

The reputation formation starts as stated above with the formation of an image. The information (experience, perceived damage) about a service/product is provided by a sender. This is called information supply.

The information diffuses in the market and reaches other customers. The diffusion speed is proportional to the degree of organization and the network density and inversely proportional to the size of the network. Another important criterion in the market is extent and currentness of the information. The more current an information is the more relevant it is for the receiver. The relevance is estimated by the receiver due to the similarity of the situation and the damage potential a transaction would bear. This means that the more similar the situation is and the higher the damage potential, the more relevant is the information for the receiver.

In order to influence the decision of a receiver the sender's information has to be considered relevant. This is only the case if there is a significant similarity in the situation and if the damage potential of the transaction is higher than the costs of acquiring information. Direct information is more relevant than second hand information because first hand information is more important than information from potentially untrustworthy recommenders. The same holds true for the credibility of the sender. If he is credible the information is more important than if the trustworthiness is unclear. The aggregation considers both credibility and relevance in order to form reputation. This is necessary because there can be oppositional recommendations for one service/product. The receiver is responsible for weighing this information. Negative recommendations can be compensated by positive ones but only to a certain degree. The above named criteria influence the information which was submitted in the market from the sender and form a reputation which influences then the receiver and potential consumer [Bü00, p. 10].

2.2 Reputation Mechanisms in On-line Environments

The mechanism of word-of-mouth described above is operationalized in multiple ways to enhance security and trust in anonymous networks such as the internet. Dellarocas states that "voluntary feedback will be underprovided" [Del05, p. 17], because if it is made available in the system, everybody can profit from it at no cost. But reputation mechanisms can ensure "cooperation and efficiency [...] without the need for costly enforcement institutions" in such environments [Del05, p. 2]. In contrast to their off-line counter parts on-line reputation mechanisms need to have unique properties which distinguish them significantly. The need is constituted in the anonymity of e.g. the internet or Multi Agent System (MAS). Feedback has usually a subjective feature and can be submitted truthfully or falsely which cannot be distinguished as it could be in off-line settings with contextual cues. The internet and e.g. distributed settings as Peer-to-Peer (P2P) networks make it additionally easy to change identities, operate under multiple identities or even manipulate by discarding a bad reputation by withdrawing from the system and starting again with a "clean" identity.

All the above described problems have to be dealt with by a reputation mechanism. Moral hazard, the temptation to defect after the other party has paid or delivered the service, is the major challenge. Therefore, the mechanism has to set the right incentives to render such a behavior irrational. Chapter 3 on the next page will introduce reputation mechanisms which try to facilitate sufficient submission of trustworthy feedback and to detect deception in the system.

2.3 Preconditions of Trustworthy Feedback

The formation of reputation involves four groups of agents (which might overlap):

Evaluators (E) or trustors[1] are usually buyers of a service or a product. They evaluate after consumption and propagate their evaluation. The agent who provides a service is called the target (T) or trustee. The beneficiaries (B) could be other market participants and future buyers and of course other evaluators. Transmitting agents (M) are usually the evaluators.

The market needs to fulfill basic requirements such as [Bü00]:

ubiquity being accessible independent of time and place

trusting agents agents who believe that the information in the market is trustworthy;

independence evaluators and targets have to be independent in order to assure unbiased feedback, preventing an exploitation of the system and rent seeking;

evaluation agents have to actually transmit feedback to other market participants to make it beneficial for other agents;

imitation agents have to imitate the other agents behavior because it is not sufficient to have only a single reputation for an agent but it is favorable to have a information from different agents.

sufficient density and organization of the network density describes the ratio of the number of direct relationships between the agents. A high density and organization allows information to diffuse quickly.

[1] We assume that only the buyers (trustors) rate the behavior of the providers (trustees), because of advanced payment. This is done for reasons of simplicity and could be the other way around, too, within the same system.

3 Implementing Incentive Setting Reputation Mechanisms in Multi Agent Systems

As we have seen in the previous chapters reputation mechanisms can establish trust in anonymous markets and MAS. The reputation mechanisms have to fulfill two main functions. They have to elicit feedback from rational agents which will not submit feedback without a framework of incentives and secondly, they have to be able to detect untrustworthy and further trustworthy feedback.

In the following we introduce five approaches that attempt to solve these two problems. The differences and advantages of each one will be presented in chapter 4 on page 31.

3.1 Liu and Issarny: An Incentive compatible Reputation Mechanism for Ubiquitous Computing Environments

Liu and Issarny [LI06] introduce a reputation mechanism which has the following objectives: It needs to be able to distinguish between trustworthy and untrustworthy agents and also between honest and dishonest recommenders. Additionally, it should achieve to enforce honest recommendations. If untrustworthy feedback is given it penalizes the dishonest behavior and punishes any exploitation of the system. Therefore only honest recommendations are taken into account. Old reputation values need to be discounted over the past because they become irrelevant when behavior of the target changes. Hence, more weight is given to recent experiences.

3.1.1 Beta Reputation

The authors use the beta distribution of reputation for modeling reputation. It expresses the probability for having an event T the next time.

The advantages of beta reputation include the simple estimation of the trustworthiness of an entity by calculating $\frac{\alpha}{\alpha+\beta}$. It is easy to calculate the number of experiences on which the estimation is depending on by calculating $\alpha + \beta - 2$. Only newcomers have a value of 0. The aggregation of observation is due to dynamic adjustment by addition and accumulation of more experiences. The time fading factor explained in section 3.1.2 allows a different emphasis on recent experiences compared to older ones.

3.1.2 Time fading

Reputation values lose relevance over time, because trustees can change their behavior. Therefore the authors introduce a time fading factor for past reputation values. The factor ρ can have any value between [0,..,1]. A low value means that past experiences are forgotten more quickly compared to a higher ρ. In extreme cases for $\rho = 0$ historic values are instantly forgotten and for $\rho = 1$ they are kept forever indicating that there is no need to discount past values.

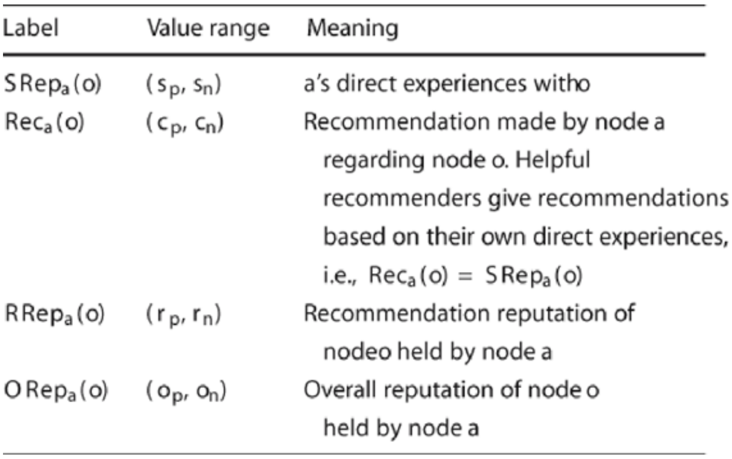
The discount formula looks like this:

$\alpha' = 1 + (\alpha - 1) \times \rho^{\Delta T}$

$\beta' = 1 + (\beta - 1) \times \rho^{\Delta T}$

3.1.3 Reputation Formation and the Three Kinds of Reputation

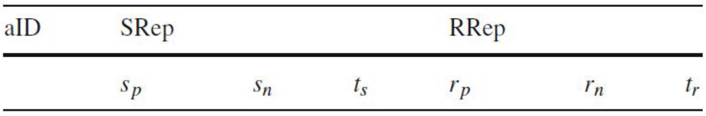
There are two roles, the trusting entity (trustor) a and a trusted entity (trustee) o. In the following the trustee will always be the provider and the trustor the consumer. Hence, $Rep_a(o)$ is o's reputation from a's point of view. The authors differentiate between three different kinds of reputation. The Service Reputation (SRep), the ReRecommendation Reputation (RRep) and the Overall Reputation (ORep). The following figure 2 lists the notations used by the authors.



Label	Value range	Meaning
$SRep_a(o)$	(s_p, s_n)	a's direct experiences witho
$Rec_a(o)$	(c_p, c_n)	Recommendation made by node a regarding node o. Helpful recommenders give recommendations based on their own direct experiences, i.e., $Rec_a(o) = SRep_a(o)$
$RRep_a(o)$	(r_p, r_n)	Recommendation reputation of nodeo held by node a
$ORep_a(o)$	(o_p, o_n)	Overall reputation of node o held by node a

Figure 2: Notations [LI06, p. 301]

Recommendation values received from other entities are stored in an acquaintance table. The *aID* is the acquaintance ID. The recommendation values are presented by two parameters, representing positive (s_p) and negative s_n experiences. t_s and t_r are the time stamps indicating when it was updated last.



aID	SRep			RRep		
	s_p	s_n	t_s	r_p	r_n	t_r

Figure 3: Acquaintance Table [LI06, p. 302]

Quality of Service (QoS) states the promised dimension of the service, e.g. an availability of 99%.

Quality of Experience (QoE) is the conformance of the advertised service and the service delivery, e.g. an advertised availability of 99% but a delivery of 80%.

The SRep combines the direct experiences one agent has with the experiences of another agent. Therefore, it is updated after each new experience. It is updated using the Quality of Experience.

Recommendation Reputation (RRep) evaluates the usefulness of a recommendation from another agent.

The Overall Reputation (ORep) describes the direct experiences an agent had from transactions if they are significant enough to derive a trust decision.

Overall Reputation

The ORep relies only on direct experiences of the trustor if those are significant enough to derive a trust decision. This is the case if the accumulation $(s_p + s_n - 2)$ reaches a certain threshold. Otherwise the trustor asks other entities for recommendations. Then the combination of own direct experiences and recommendations from others makes up the ORep of the trustee (e.g. entity o). An example will clarify this. Entity a asks entity r for recommendations about o. Then r gives $Rec_r(o) = (r_p, r_n)$. a checks then if the recommendation is trustworthy in two steps. (1) Is r honest? If $\frac{r_p}{r_n+r_n}$ is high enough, r is considered honest. (2) The RRep is evaluated with $(r_p + r_n - 2)$ to ensure it relies on enough evidences. If those two criteria can be met by the recommendation of r, the recommendation is taken into account and weighted according to the formula:

$$w_r = E(Beta(r_p, r_n)) = \frac{r_p}{r_p + r_n}$$

This is done for each recommendation. The complete ORep is then calculated from the sum of all those:

$ORep = \delta \times SRep + (1 - \delta) \times \frac{\sum_{r \in R}(Rec_r(o) \times w_r)}{\sum r \in R(w_r)}$.

The δ represents the weight given to each recommendation. It is usually greater than 0.5 due to the fact that own direct experiences are more valuable than recommendations from other entities. ORep is not kept as an acquaintance record but is dynamically evaluated when needed since it evolves over time with new experiences added when possible.

Quality of Service and Quality of Experience

In order to identify untrustworthy providers the experience is described with the metric of Quality of Experience (QoE). The providers advertise their Quality of Service (QoS) which could be availability, delivery at a certain time, etc. After the transaction has occurred and is finished, the consumer can rate the conformity of the QoS with the QoE. The QoS consists of the dimensions d_i (i = 1,...,n), e.g. availability, latency etc. The promised value has the form p_p (i = 1,...,n). The consumer receives a quality stated by a_i (i = 1,....,n), this is the actual value for the promises p. The assessment of the specific quality of service consumer a has experienced with provider p is done with the following formula:

$$QoE_a(o) = \sum_{1 \leq i \leq n} \frac{comp(a_i, p_i)}{n}$$

$comp(a_i, p_i)$ is the function to calculate the degree of conformance for one dimension, e.g. availability, between the actual a and the promised p QoS.

Now we want to look at three different cases in which values are inserted and the actual QoS is calculated. First, we assume a simple promise. The provider promised to deliver. A a of 1 would imply that the request was satisfied as advertised, 0 that it was not. The comp function would look like this if the request was satisfied or not:

$$comp(a_i, p_i) = MIN(1, \frac{a_i}{p_i})$$

The comp function would yield $comp(1,1)$ if it was satisfied, and $comp(1,0)$ respectively if it was not. Considering the case of availability: The operators in the comp-function look like the following because the dimension is stronger with larger values. This means that a larger value is better than a smaller one.

For p = 98% , a = 100%:

$comp(a_i, p_i) = MIN(1, \frac{1}{0.98})$.

In the case of latency is stronger with smaller values:

$comp(a_i, p_i) = MIN(1, \frac{p_i}{a_i})$.

For p = 0.8ms , a = 1ms: $comp(p_i, a_i) = MIN(1, \frac{0.8}{1})$.

The addition of the single comp-functions when more than one dimension is advertised returns the overall QoE.[2]

$$QoE = \frac{MIN(1, \frac{0.8}{1.0}) + MIN(1, \frac{1.0}{0.99})}{2} = 0.9$$

[2] Liu et al. also consider the case of a dimension i with Boolean values. We will not deal with those for reasons of simplicity.

The QoE is used to update the SRep

1. $s'_p = s_p + QoE$
2. $s'_n = s_n + (1 - QoE)$

Recommendation Reputation

The RRep is exclusively made up of direct experiences of using recommendations. It has the form (c_p, c_n) and is equal to the SRep for honest recommenders. "Given a new QoE of $e \in [0...1]$ the honesty of a recommender is adjusted according to the helpfulness of its recommendation" [LI06, p. 303]. Liu et al. provide a calculation for this which is beyond the scope of this paper for further information see [LI06, p. 303].

The beta reputation (see figure 4) provides now a simple calculation to check whether an agent is an active recommender: $r_p + r_n - 2$. The value is expected to be high for active recommenders. To check whether an agent is providing honest recommendations the value of $f(p|r_p, r_n)$ is expected to be high, too. As the following graphs show. The higher the first value (r_p) is the more positive values were observed. The higher the sum the higher is the number of recommendations the agent has given.

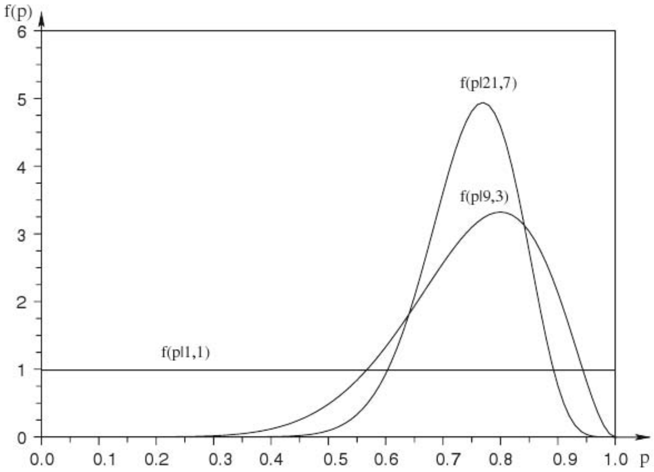


Figure 4: Beta distribution for RRec [LI06, p. 300]

The two values δ_h and δ_a are the thresholds for trustworthiness (honesty) and activeness in providing recommendations. Therefore a provider would be considered active if $r_p +$

$r_n - 2 \geq \delta_a$ and honest if $\frac{r_p}{r_p+r_n} \geq \delta_h$. This leads to five distinct states of a recommender: active truthteller, inactive truthteller, active liar, inactive liar and newcomer.

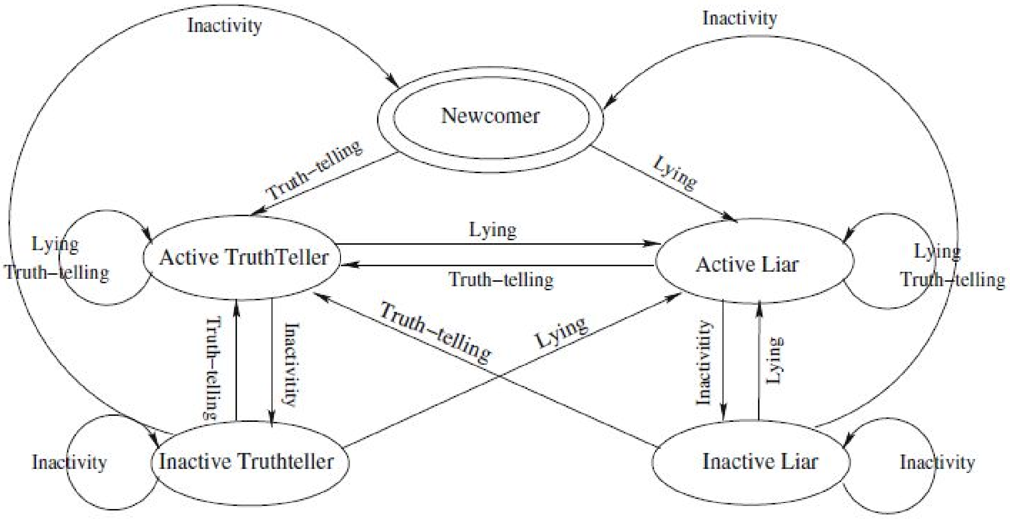


Figure 5: States of the Recommenders [LI06, p. 304]

The different states of a recommender change due to behavior, activity and inactivity.[3] RRep decays if an agent does not provide recommendations and moves him from an active liar or truthteller to an inactive counterpart or even a newcomer. The distinction between five groups of recommenders is crucial for the reputation propagation because the groups are treated differently in granting access to reputation information. Hence, these five states set incentives to share honest recommendations with other agents.

3.1.4 Incentives in the Reputation Propagation

As seen above, the RRep is exclusively made up of direct experiences from recommendations. It takes only recommendations from truth tellers into account. If there are none, the average of the recommendations from inactive truthtellers and first time recommenders is calculated. Then the trustee's ORep is calculated according to the formula from section 3.1.3 on page 10. Otherwise he has to rely on his own direct experiences which might be too few in a MAS because there are many participants and a high fluctuation of members in the market. After the service consumption the QoE is updated (see section 3.1.3 on the preceding page) and compared to all other recommendations in order to update

[3] The loop e.g. at Active truth teller indicates that lying worsens the reputation but does not destroy it immediately. Only multiple lies make an active truthteller an active liar.

the recommenders' RRep. Now liars can be identified and their RRep is updated, too. Since all recommendations were accepted before, but only the honest ones were taken into account the agent can now compare the recommendations from the classified liars to the outcome. This gives them the chance to improve or worsen their RRep.

If an agent o then asks agent a for recommendations, agent a first evaluates the state of the agent o and if he himself has a significant number of direct experiences. If he does and agent o is an honest recommender he sends back the recommendation immediately. In the case that agent o is considered inactive he sends back the recommendation with the probability of $diff = \delta_a - (r_p + r_n - 2)$. The distinction between inactive liars and truthteller is made by the fact that inactive recommenders do not necessarily withhold their recommendations. They are treated differently by changing the formula by a small value of ϵ (decreasing for liars and increasing for truthtellers). "Therefore the less active an entity is, the less possible that it receives helpful recommendations from others" [LI06, p. 304].

3.2 Jøsang and Ismail: The Beta Reputation System

The reputation system introduced by Jøsang and Ismail [JI02] is based on the beta probability function which reflects the probability distribution of binary events. Unlike Liu et al., Jøsang et al. use a centralized setting with a collection center to store reputation values, because they intended it for human actors in e-commerce environments. Their reputation mechanism can also be used in distributed settings such as the MAS we are looking at.

3.2.1 Beta Density Function and the Reputation Formation

The beta density function for reputation represented looks slightly more complicated than the one presented by Liu et al., but ends up to be similar. They use the gamma function Γ:[4]

$$f(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \text{where } 0 \leq p \leq 1, \alpha < 0, \beta < 0,$$

[4] which is similar to the beta function but includes additionally complex and real numbers

The expectation value within the restrictions $p \neq 0$ if $\alpha < 1$and $p \neq 1$ if $\beta > 0$ is similar to Liu et al. again:

$$E(p) = \frac{\alpha}{(\alpha + \beta)}$$

There are again two possible outcomes, here called x and $\overline{x}$ which are corresponding to T and $-T$ in Liu et al.'s beta reputation. The observed number of x is called r and of $\overline{x}$ is s both of them need to be greater or equal to zero $(r, s \geq 0)$. The probability density function of observing outcome x in the future can be expressed as a function of past observations by setting:

$$\alpha = r + 1$$

and

$$\beta = s + 1$$

where

$$r, s \geq 0$$

. With the beta function they are trying to visualize that the relative frequency of outcome x in the future is somewhat uncertain and that the most likely value corresponds to $E(p)$. Hence, the reputation function predicts the expected relative frequency with which x will happen in the future. The authors use super- and subscript to indicate the provider (superscript) and the target (subscript). Hence, (r_T^X,s_T^X) represents the positive and negative feedback tuple about target provided by entity X. Those tuples are called reputation parameters. The probability expectation value of reputation function is accordingly.

$$E(\varphi | p(r_T^X, s_T^X)) = \frac{r_T^X + 1}{r_T^X + s_T^X + 2}.$$

This again is similar to Liu et al. where it is defined as $E(p) = \frac{\alpha}{(\alpha+\beta)}$. Jøsang et al. add that their model does not provide objectivity because honesty cannot be enforced with this reputation mechanism which is also true for Liu et al. but is treated differently because Liu et al. make use of RRep to enforce honesty in rational agents.

3.2.2 Reputation Rating and Combining Feedback

In the reputation rating and representation, Jøsang et al. make an important distinction. As mentioned above, their reputation mechanism targets e-commerce participants especially human actors and not so much rational agents. Therefore they introduce a reputation rating between [-1,+1], with 0 as a neutral value. The probability expectation representation with E(p) is very suitable but unfamiliar and confusing for most human users. Therefore they introduce a single feedback value which will not be specified any further at this point because it is not necessary for rational agents to simplify the probability functions. The accumulation of feedback is similar to Liu et al. again. When feedback from entity X (r_T^X, s_T^X) and entity Y (r_T^Y, s_T^Y) about target T is received the r-parameters and the s-parameters are added up as follows:

$$r_T^{X,Y} = r_T^X + r_T^Y \text{ and } s_T^{X,Y} = s_T^X + s_T^Y$$

This leads to the updated reputation function $E(\varphi|p(r_T^{X,Y}, s_T^{X,Y}))$. Jøsang et al. add that the independence between the ratings must be assumed so that no feedback can count twice.

3.2.3 Discounting

Belief Discounting

The authors present two different kinds of discounting. First belief discounting because "feedback from highly reputed agents should carry more weight than feedback from agents with low reputation rating" [JI02, p. 6]. Therefore, they introduce w_T^A which reflects the opinion of A about target T. The opinion consists of belief, disbelief and uncertainty:

$w_T^A = (b, d, u)$ where $b + d + u = 1$ and $b, d, u \in [0, 1]$
b = probability that proposition x is true; $= \frac{r}{r+s+2}$
d = probability that proposition x is false; $= \frac{s}{r+s+2}$
u = mass that is unaccounted for; $= \frac{2}{r+s+2}$

In belief discounting an advice from Y to X about T is expressed as $w_T^Y = (b_T^Y, d_T^Y, u_T^Y)$. Now the advised X rates the opinion and comes to a derived opinion about T. X's opinion about T as a result of Y's advice to X is represented as:

$w_T^{X:Y} = (b_T^{X:Y}, d_T^{X:Y}, u_T^{X:Y})$

This function depends on b, d and u as defined above.

$b_T^{X:Y} = b_Y^X b_T^Y$

$b_T^{X:Y}$ means that agent X discounted the advice from Y about T by its opinion about Y. This is also done for d and u.

$d_T^{X:Y} = d_Y^X d_T^Y$

$u_T^{X:Y} = d_Y^X + u_Y^X + b_Y^X u_T^Y$

After belief discounting the authors introduce reputation discounting in order to discount "feedback as a function of the reputation of the agent who provided the feedback" [JI02, p. 6]. The incentives set by Jøsang et al. are similar to Liu et al. [LI06] because they both establish a "meta-rating" reflecting an indication how truthful the agent reports. Liu et al. call it recommendation reputation (RRep) and Jøsang et al. call it belief.

Reputation Discounting

Secondly, they introduce the reputation discounting which is slightly different from the discounting methods used before. The authors take into account that a recommendation must not necessarily be true and consider the opinion the agent has about the target and the recommender. $\varphi(p|r_T^{X:Y}, s_T^{X:Y})$ is the reputation function of T given a recommendation from Y which is discounted by agent X. This means that the given function is T's discounted reputation function by X through Y.

$$r_T^{X:Y} = \frac{2r_Y^X r_T^Y}{(s_Y^X + 2)(r_T^Y + s_T^Y + 2) + 2r_Y^X}$$

$$s_T^{X:Y} = \frac{2r_Y^X s_T^Y}{(s_Y^X + 2)(r_T^Y + s_T^Y + 2) + 2r_Y^X}$$

Forgetting

Similar to Liu et al. [LI06] Jøsang et al. introduce a forgetting factor which discounts old feedback in order to adapt to behavior changes of the ratee. Hence, old feedback is given less weight than more recent feedback. This corresponds to “gradually forgetting” [JI02, p. 7] feedback values. The authors take a sequence of feedback values and show how it is discounted. The order in which the feedback is observed is very important because of the different weights of the single values. The disadvantage is that all feedback has to be stored forever which can lead to capacity shortages.

3.3 Buchegger and Boudec: A Robust System for P2P and Mobile Ad-hoc Networks

Buchegger and Boudec [BB04] create a reputation system which detects misbehavior but does not set any direct incentives to submit reputation. The only incentives set are used to enforce correct feedback and to maintain a good personal reputation. In order to create incentives for rational agents to submit feedback at all more mechanisms have to be implemented as proposed by Liu et al. Where reputation values are not or only given with a slight chance to other participants who do not appear to be an active truth teller (section 3.1.4 on page 13). Although Buchegger et al. lack this important feature for our setting they still propose a very interesting reputation mechanism which is fully distributed and does not require any central institution or agreement beforehand. Like the other reputation mechanisms introduced so far it uses Bayesian estimation to detect false reports.

3.3.1 Reputation Representation

The reputation of a given agent (which the authors call node) is the collection of ratings about this agent. This information is kept and maintained by others instead of being stored in a centralized institution. Hence, the reputation system is fully distributed. Reputation values appear in three different kinds. First of all the reputation rating ($R_{i,j}$) which indicates the opinion of agents i about agent j’s behavior in the system. The trust rating ($S_{i,j}$) expresses agent i’s opinion about how honest agent j is. Those two ratings and additionally the first hand information ($F_{i,j}$) from agent i about agent j make up the reputation of agent j maintained by agent i. The three kinds of reputation values are represented in tuples so that e.g. $F_{i,j}$ has the parameters (α, β) of the Beta distribution by agent i in its Bayesian view of agent j's behavior, initially set to (1,1).

Reputation Building, Updating and Discounting

When a agent i makes a first hand experience with agent j it updates $F_{i,j}$ and $R_{i,j}$ so the first hand experience rating and the rating about agent j's behavior in the base system. From time to time the first hand ratings are published and participants can include them in their reputation ratings about other agents. In order to integrate the published rating agent i has to estimate if the other agent, here agent k, is trustworthy.

If agent k is considered trustworthy or the submitted $F_{k,j}$ is close to $R_{i,j}$ the first hand information $F_{k,j}$ is accepted and used to slightly modify $R_{i,j}$. If it does not satisfy one of these criteria the $R_{i,j}$ is not updated. In every case the trust rating $T_{i,k}$ is updated which is similar to Liu's approach but does not go as far. The trust rating slightly improves if $F_{k,j}$ is close to $R_{i,j}$ or slightly worsens if not. It then helps to maintain an opinion about the honesty of a agent.

At this point Liu et al. introduced the possibility to categorize the agents as active truthteller, active liar, the inactive counter parts and newcomers. This would be suitable for Buchegger et al., too, because it would enable to distinguish those different kinds of agents when publishing $R_{i,j}$ in the future and holding information back in order to punish not submitting feedback at all and misbehavior.

During the publication process only $F_{i,j}$ is submitted; $T_{i,j}$ and $R_{i,j}$ are never disseminated. Updating the reputation of agent j by direct experiences, agent i, works as follows. The observation made by an agent can have the form of s = 1 for misbehavior or s = 0 otherwise. Hence, s is defined as $s \in [0, 1]$ Then the new reputation value is computed by

$$\alpha := u\alpha + s$$

$$\beta := u\beta + (1 - s)$$

with u being the discount factor in order to enable "forgetting" or more technically reputation fading due to time because the agents can change their behavior over time.

This approach is quite similar to Liu et al. [LI06] and Jøsang et al. [JI02] but not as advanced because it only allows binary results such as delivery successful yes/no. What differentiates Buchegger et al. [BB04] from the others is that they give a method to find out a good value for the discount factor u.

3.3.2 A Good Value for u

In a sequence of observations $s_1, ..., s_n$, more weight should be given to more recent observations. In order to allow time fading, a good value of u has to be estimated. u should be greater than zero but not greater than one ($0 < u < 1$), so that fading is enabled and lesser weight is given to older observations. From the equation for the new α after an observation we can derive a standard formula for n n observations so that

$$\alpha_n = s_n + us_{n-1} + ... + u^{n-1}s_1 + u^n$$

This series shows that each observation is less weighted when a new observation is made and that without any observations the α-value still fades (expressed with the last u^n).

In order to find a good value for u they introduce θ as the probability that agent j misbehaves in a transaction with agent i. To compute the expected value of α after a large number of n observations (α_n) they assume θ to be constant

$$E(\alpha_n) \approx \frac{\theta}{1-u} \text{ and } E(\beta_n) \approx \frac{1-\theta}{1-u}$$

respectively for the expected value of β after a large number of n observations. Then they introduce a m as an integer with $m = \frac{1}{1-u}$. This m additionally represents the number of observations in which stationary behavior of the other agent can be assumed. So that

$$u = 1 - \frac{1}{m}$$

This makes the discount dependent on the behavior volatility of the other agent which makes perfectly sense when considering the extreme cases of a behavior change every time the agent enacts. Here m would equal 1 so that $u = 1 - \frac{1}{1} = 0$. This means that the old experience is useless for estimating the probability that agent j will defect the next time. In the case that agent j only changes it behavior after ten observations u = 0.9 which gives older observations still a pretty high weight and time fading is much slower.

3.3.3 Trust Ratings

Trust ratings help the agents to estimate how honest another agent is. They are updated whenever a report about an agent is published. The process works as follows: Agent i believes that every other agent provides false reports with a certain probability. Let the probability of agent k providing false reports be ϕ. In order to estimate the expectation of the distribution of ϕ agent i uses the prior $Beta(\gamma, \delta)$. The trustrating $T_{i,j}$ is therefore equal to (γ, δ). This is set initially to (1,1). In order to test a rating the deviation test introduced in section 3.3.4 is used whether the agent k is already considered trustworthy or not. If the deviation test succeeds $s = 1$, $s = 0$ if not. After the test the trust rating is updated very similar to the updating before with a discount factor v:[5]

$$\gamma := v\gamma + s$$

$$\delta := v\delta + (1 - s)$$

3.3.4 Reputation Rating and Model Merging

Similar to the first hand observations the reputation ratings have the form $R_{i,j}$ which has the parameters (α', β'), initially set to (1,1). $R_{i,j}$ is always updated when a first hand observation is made ($F_{i,j}$ is updated) and when $R_{k,j}$ from another agent is published and accepted. The update due to a new $F_{i,j}$ functions just like updating $F_{i,j}$ so that

$$\alpha' := u\alpha' + s$$

[5] The factor v is similar to u but since it does not necessarily have the same values as u it is called v.

$$\beta' := u\beta' + (1 - s)$$

An inactivity update, in order to enable time fading just removes the last part of the two equations: $\alpha' := u\alpha'$ and $\beta := u\beta' + (1 - s)$.

If agent i receives a first hand observation $F_{k,j}$ from agent k about agent j, agent i tries to find out if this information is correct by taking trust and compatibility into account. Agent i will then check if agent k reaches the threshold for honest recommendations (defined below) if it does it will include $F_{k,j}$ in $R_{i,j}$ as follows.

$F_{k,j}$ is modified by a factor w which is a small positive constant that allows agent i to give the feedback from agent k a different weight than its own reputation ratings. $F_{k,j}$ is then added to $R_{i,j}$: $R_{i,j} := R_{i,j} + wF_{k,j}$.

If agent k is considered untrustworthy it will apply a deviation test. $E(Beta(\alpha, \beta))$ is defined as the expectation of the distribution $Beta(\alpha, \beta)$. What they do then is to compare the expectations of the distribution of the tuples from $F_{k,j}$ and $R_{i,j}$ if they reach a certain threshold d. Here $F_{k,j}$ has the parameters (α_F, β_F) and $R_{i,j}$ the parameters (α, β):

$$|E(Beta(\alpha_F, \beta_F)) - E(Beta(\alpha, \beta))| \geq d$$

If the deviation test is positive, agent i will not consider the first hand information $F_{k,j}$ because it is incompatible. Otherwise $F_{k,j}$ is used to update $R_{i,j}$ as if $F_{k,j}$ would have been considered trustworthy.

3.3.5 Decision-Making Process

At first all the information from first hand experiences is taken into account which means that all $R_{i,j}$ and $T_{i,j}$ are updated. To make a final decision the beta distribution is once again used. This is similar to the method used for the reputation rating. The first estimation is done for $R_{i,j} = (\alpha', \beta')$. They consider $E(Beta(\alpha', \beta'))$ for θ so that normal behavior would satisfy:

$$E(Beta(\alpha', \beta')) < r$$

Misbehaving would be indicated when $E(Beta(\alpha', \beta')) \geq r$. The same is done for $T_{i,j} = (\gamma, \delta)$. It is considered trustworthy for:

$$E(Beta(\gamma, \delta)) < t$$

In the case of $E(Beta(\gamma, \delta)) \geq t$ agent i would consider agent j as untrustworthy. “The thresholds r and t are an expression of tolerance.” Therefore $r = 0.5$ would imply that misbehavior in less than half of the times. Similar to that $t = 0.75$ implies that lying in less than 25% of the cases is tolerated.

3.3.6 Incentives

Buchegger et al. use just like Liu et al. [LI06] a trust rating that estimates how truthful another agent reports. This “meta rating” allows different treatment if the other agent asks for feedback or when their feedback is incorporated for decision purposes. Liu et al. go a step further than Buchegger et al. in this point, because he does not automatically publish recommendations but evaluates the other agent and sends a recommendation only with a certain probability back, according to the state of the asking agent (see section 3.1.3 on page 12). According to Buchegger's approach, recommendations are published automatically and all agents have access. The difference here is that they introduce a factor u which allows an estimation how long the agents behavior is stable and whether it can be trusted. That makes the stored recommendations (direct and indirect) more valuable and therefore it sets an incentive to behave accordingly to the factor u.

3.4 Yu and Singh: A Social Mechanism of Reputation Management in Electronic Communities

Yu and Singh [YS00] have developed a social reputation mechanism that tries to avoid interactions with untrustworthy agents. The mechanism is social because the agents trade feedback about possible interaction partners and gossip is additionally used as a source of information. The agents are assigned a unique ID which makes them distinct from others in the system. In order to gain information, agents pose queries to the system and wait for others to respond to them. The queries always include the question, the ID and the address of the agent, additionally a limit of the referrals requested is given. The other agents can then decide either to answer the query after assessing if it has enough information and/or to give a referral. A referral is only sent if the questioning agent is trusted by the answering agent. After receiving the response from the other agent the originating agent weighs the answer and updates its "opinion" about the answering agent. This is important because Yu et al. describe the questioned agents as neighbors and agents are interested in questioning only reliable sources. A referral from the answering agent is judged, too, and then decide whether to rely on it or not.

3.4.1 Reputation: Referral Chains and Gossip

Yu et al. [YS00] distinguish two different methods of acquiring feedback. First of all it can be acquired through the above described referral chains which means that if agent A trusts agent B, and B trusts C, then A is more likely to trust C as well. Second, gossip is treated differently because "an agent can propagate a rumor without having been explicitly queried." Therefore "gossip is processed incrementally". [YS00, p. 6]. So the trust rating T has to be treated differently. $T_i(j)^t$ is defined as the trust rating agent i has about agent j at time t. A positive evidence increases the trust rating by α ($\alpha > 0$) and a negative evidence decreases it by β ($\beta < 0$). In order to punish undesired behavior Yu et al. set $|\alpha| < |\beta|$ which enables reputation to tear down easily but being hard to build up.

3.4.2 Incorporating Feedback

Trust ratings are changed when a direct observation occurs (due to an interaction), a feedback from another trusted agent is received, gossip is submitted into the system. The authors distinguish three ways to acquire information to update the trust ratings: The direct interaction, the testimony from another witness and gossip.

Direct Experiences

There are six cases in case the agents have interacted before. The following table shows in the first column how agent i and agent j have interacted before. Across you find the ongoing transaction behavior of agent j (cooperate or defect).

Table 1: Incorporating trust from direct experiences

In past transactions agent i... agent j	cooperates	defects
...trusted	$T_i(j)^{t+1} = T_i(j)^t + \alpha(1 - T_i(j)^t)$	$T_i(j)^t + 1 = \frac{T_i(j)^t - \beta}{1 - min(\lvert T_i(j)^t\rvert, \lvert\alpha\rvert)}$
...did not trust	$T_i(j)^{t+1} = \frac{T_i(j)^t + \alpha}{1 - min(\lvert T_i(j)^t\rvert, \lvert\alpha\rvert)}$	$T_i(j)^t + 1 = T_i(j)^t + \beta(1 - T_i(j)^t)$
...did not interact before with	$T_i(j)^{t+1} = \alpha$	$T_i(j)^t + 1 = \beta$

Testimonies from Other Witnesses

Testimonies from other witnesses are not directly incorporated. First the testimonies from witnesses who are not considered trustworthy are discarded. If there is more than one testimony from one witness only the best one is considered. Then a mean is calculated from the remaining testimonies, this is called $\bar{E}$. There are four possible cases:

Table 2: Incorporating testimonies

In past transactions agent i ... agent j	*then*
...trusted and the witnesses trusted	$T_i(j)^{t+1} = T_i(j)^t + \bar{E}(1 - T_i(j)^t)$
...trusted but the witnesses did not trust	$T_i(j)^{t+1} = T_i(j)^t + \frac{\bar{E}}{1 - min(\lvert T_i(j)^t\rvert, \lvert\bar{E}\rvert)}$
...did not trust but the witnesses trusted	
...did not trust and the witnesses neither	$T_i(j)^{t+1} = T_i(j)^t + \bar{E}(1 + T_i(j)^t)$

Gossip

The third way to acquire information about agent j is taking gossip ($T_i(j)$ from agent k about agent j into account. Again there are four cases:

Table 3: Incorporating Gossip

In past transactions agent i...	*then*
...trusted both agent j and agent k	$T_i(j)^t + 1 = T_i(j)^t + T_i(k)^t \cdot T_k(j) \cdot (1 - T_i(j)^t)$
...neither trusted agent j nor agent k	$T_i(j)^t + 1 = T_i(j)^t + T_i(k)^t \cdot T_k(j) \cdot (1 + T_i(j)^t)$
...did not trust agent j but trusts agent k	$T_i(j)^t + 1 = \frac{T_i(j)^t + T_i(k)^t \cdot T_k(j)}{1 - min(\lvert T_i(j)^t \rvert, \lvert T_i(k)^t \cdot T_k(j) \rvert}$
...trusted agent j but does not trust agent k	

3.4.3 Incentives

Yu et al. [YS00] do not provide direct incentives for rational agents to report and to do that truthfully but they create a reputation mechanism in which trust building is very hard and tearing down the trust rating is fairly easy. Therefore agents will try to misbehave as much as they can without damaging their trust rating. To prevent such behavior the authors give an example. θ is the ratio between the times an agent cooperates and defects. If the ratings for α and β are appropriately selected $\theta \to \infty$. Additionally, gossip increases the diffusion of information among all agents in the system even if they have not interacted before.

3.5 Jurca and Faltings: Towards Incentive Compatible Reputation Management

The reputation mechanism represented by Jurca and Faltings [JF03] introduce a mechanism to detect false feedback and additionally a framework of incentives which make it rational to report truthfully for rational agents. They do that by introducing a side payment scheme which is maintained by broker agents. Those are called Broker Agent (R-Agent)s and they are the only ones who can trade with reputation values.

The following assumptions are made by Jurca et al.:

1. Payments are only conducted by R-Agents. No side payments occur between any normal agents.

2. All agents behave rationally.

3. There are n agents in the system with a_i for $i = 1...N$.

4. Agents play in pairs iterated prisoner's dilemmas.

3.5.1 Acquiring Feedback from Other Agents

As we have seen in section 2.2 on page 5, it is useful not only to rely on direct experiences but to acquire further feedback from other agents in the system as well. Therefore it is rational for agents to try to acquire information from other agents. In this mechanism they can do that by buying information about another agent at the cost of F from an R-Agent. After the transaction, given that it has taken place, the agent can sell reputation information for C. The optimal value for N and C will be estimated in section 3.5.4 on page 29. Agents are only allowed to sell reputation to an R-Agent about an agent that they have purchased information about before. The agents buy systematically reputation information before interacting with another agent in Jurca and Falting's scenario.

3.5.2 Reputation Information

In contrast to the reputation mechanisms introduced so far, Jurca et al. use a single real number representation of the reputation information r_i. It can have the value 0 for defecting and 1 for cooperating behavior. Hence, the reputation lacks the accuracy that Liu et al. and the others have by introducing a span for reputation information as $r_i \in \{0, 1\}$. However, in later articles Jurca et al. show that their model works with other values, too [JF07, JF06]. Reputation can be calculated by:

$$r_i = frac \sum_{j=1}^{k} report_j N$$

So that the reputation value is computed as the average of all the reports about that specific agent. The $report_j$, $j = 1...k$ represents all the reports for that agent a_i.

3.5.3 Incentive Compatibility

In order to make the mechanism incentive compatible, the following features of the model are assumed by the authors:

1. Agents which report truthfully at all times should not lose any money as a result of an interaction with another agent:
   $E[F] \leq E[C|\text{truthful report}]$

2. Agents who do not report truthfully should gradually lose their money as a result of an interaction with another agent:
   $E[F] \geq E[C|\text{false report}]$

R-Agents will pay only for reports which match the next report about the concerned agent. This is done because — as we will see below — it is optimal for a rational agent to report truthfully because he will be paid at least 50% of the cases. This was calculated by consideration of the probabilities of different behavior schemes:

- agent a_i cooperates in two consecutive rounds: p_i^2
- agent a_i defects in two consecutive rounds: $(1-p_i)^2$
- agent a_i cooperates then defects: $p_i(1-p_i)$
- agent a_i defects then cooperates: $p_i(1-p_i)$

This means that the probability of acting in the same way in two consecutive rounds is:

$$(1-p_i)^2 + p_i^2 <=> 1 - 2p_i + 2p_i^2$$

which is bound by [0.5,1]. The probability for a change in behavior in two consecutive rounds is: $2p_i(1-p_i)$ which is bound by [0,0.5]. Then Jurca et al. assume that other agents report the truth and that a_i will behave in the same way in the next round. Hence, it is rational for the agent to report truthfully because he is paid with a probability of not less than 0.5. Those assumptions are slightly different than the ones made by Buchegger et al. [BB04] because they introduce a function that estimates a time span in which the agent beliefs that the transaction partner acts the same way over multiple rounds. The assumption that the behavior is the same in consecutive rounds is needed for the calculation of the payoff by Jurca et al. Therefore their mechanism is more static. In later works they have eliminated this assumption and created a more flexible mechanism (see [Jur07a, Jur07b, JF08].)

3.5.4 Payments for Reputation

Agents purchase information about a possible prospective interaction partner but can only sell information if they did interact with that agent. Business only takes place if both agents agree. Hence, the agent can expect a payoff after analyzing three possible situations:

1. When the reputation of a_i that the agent purchased from an R-Agent is too low he will not interact with that agent and can therefore not sell any information. The payoff is 0;
2. when business has taken place and he submits a report to a R-Agent but it is considered false because the other agent has changed his behavior in the next round. The payoff is 0;
3. when business is conducted and the other agent behaves accordingly in the next round, the payoff equals C.

The expected payoff can be computed as follows:

$$E[\text{payoff}] = 0 \cdot Pr(\text{case } 1) + 0 \cdot Pr(\text{case } 2) + C \cdot Pr(\text{case } 3)$$

As stated above agents only interact with other agents if they expect a profit. This means that the probability that an agent will trust and interact with another agent q is equal to the probability of a positive outcome Out: $q = Prob(Out > 0)$.

$$Out = \frac{1}{2}[(1 - p_i) \cdot f(\frac{I}{2}) + p_i \cdot f(I)] - \frac{I}{2}$$

This is the business payoff function when I units have been invested. In this function $Out > 0$ is equal to $p_i > \theta$ if a monotone increasing function is assumed. θ is a constant that the authors use which depends only on the business payoff function. The constant is used to define q which equals the probability that p_i is greater than θ: $q = Pr(p_i > \theta)$

In order to estimate the payoff now, we need the probability that the agents interact with another. The probabilities for case 1 and 2 are given by Jurca et al. but are not considered here to make simpler because the payoff would be 0 in those two cases (for explanation see above).

$$Pr(case \quad 3) = q^2(1 - 2p_j + 2p_j^2)$$

So that the average value of the payoff and therefore the price is:

$$E[payoff] = C \cdot \frac{\sum Nj = 1q^2(1 - 2p + 2p_j^2}{N} = F$$

With this function we can compute the average payoff for the seller and the price for the buyer (F) with the help of the payments made to acquire the reputation (C).

4 Conclusion

This paper has shown that reputation mechanisms can establish trust in anonymous markets and MAS. We have looked at the specifics of agents and MAS which enable fast transactions without human interference and human supervision. This entails problems, such as a lack of trust within the system and incentives to cheat which will lead to a collapse of the system. In this respect, the prisoner's dilemma exemplified the problem with self-interested rational agents. Thereupon a framework of incentives has to be created in order to realize a equilibrium in an iterated prisoner's dilemma in which all agents cooperate. The five introduced reputation mechanism tried to do that by using different methods to elicit reputation and with by setting incentives differently. We have decided on presenting not only the incentive structure the mechanisms provide but the whole process of feedback elicitation, processing, storing and using the information and the detection of false feedback as well in order to understand the mechanisms wholly and to be able to understand why incentives are set in the particular way they do.

The five reputation mechanisms presented in chapter 3 can be distinguished by their ability to elicit honest recommendations and how they set incentives to provide feedback.

In 3.1 on page 7 we discussed *Liu and Issarny's* approach to work with three different reputation values ORep, RRep and SRep in order to estimate the trustworthiness of an agent. We have seen that the agents rate their partners due to the reputation values and divide them into the groups (called states of the recommender) active truthteller, inactive truthteller, active liar, inactive liar and newcomer. This represents incentives because if an agent sends an request for information (second hand recommendation) it is given an answer due to his state. Therefore, all rational agents will try to become a truthteller which is active in order to receive the most answers to his requests.

Jøsang and Ismail in 3.2 on page 14 is introduced in this paper because it is fairly easy to implement but still rests on a sound statistical basis. Additionally, the authors present three different discounting methods (Reputation discounting, belief discounting and forgetting) that give a detailed approach how to rate feedback from other agents. In order to estimate the trustworthiness of feedback from an agent the three factors belief, disbelief and uncertainty are taken into account and are weighted with the opinion the agent has about the feedback provider and the target agent. The mechanism does not set any further incentives than that and has to be modified further to be fully satisfying for a MAS with rational agents.

We have shown in 3.3 on page 18 that *Buchegger and Boudec* also introduce a reputation rating and a trust rating. This reputation mechanism publishes information regularly inside the system and is accessible for all participants. Their incentives are not as clear cut as with the other authors but a close examination shows that Buchegger uses a very precise estimation on how honest the provider of information is and can therefore detect false reports very quickly and refrain from conducting business with the concerning agent. The estimation how honest the provider of feedback is carried out by calculating a factor u that estimates after how many times the agent changes its behavior from cooperation to defection and back again.

In 3.4 on page 23 we presented that *Yu and Singh* provide a special feature: they distinguish reports from other agents in testimonies and gossip according to how good they know each other. Therefore, there is a different incorporation of feedback depending on how close the agents are related and on the experience the agent had with the trustee before. This leads to a very diversified incorporation of feedback ensuring that is optimally valued due to trustworthiness. This again is not a direct incentive but enforces honest behavior without the necessity of setting incentives for reputation propagation. Since their mechanism is a social one they assume that there are close related agents called neighbors which share information on a regular basis and further related agents whose feedback is considered as gossip.

Finally in 3.5 on page 26, we show *Jurca and Faltings* achieve to incorporate both: elicitation of honest feedback and setting incentives to provide feedback by payments. The authors introduce R-Agent which are broker agents. They serve as a mediator who collect feedback and sell it to other agents. After the transaction with the agent which the other agent had bought feedback about, he can sell that information to th R-Agent again. Additionally, the submitted reports are checked if they are honest or not and only paid fro if they appear honest.

The table 4 summarizes the main characteristics of each of the reputation mechanisms.

In conclusion, we would like to say that each of the five approaches stresses a very important aspect which should be considered in a “perfect” reputation mechanism. From Liu and Issarny we have to take into account the three ratings and the ability to estimate differences in the advertises and delivered service. From Jøsang and Ismail we would incorporate the three different kinds of discounting feedback in order to rate feedback precisely according to the trustworthiness of the recommender and our opinion about the target agent. Buchegger and Boudec would contribute the factor u that allows an estimation of how stable a target agents behavior is. This is important for discounting of feedback and taking behavior changes into account. Yu and Singh can bring in a social component if necessary by distinguishing closer related and distant agents. Finally, Jurca and Faltings provide the incentive setting payment mechanism that rewards submission of feedback. By combining the strengths of all the approaches, one could design a reputation mechanism that elicits feedback successfully and eliminates untrustworthy behavior through a very precise detection of it.

Further work has been carried out by all the authors who have been considered in this paper. Especially Jurca has published papers going beyond the ideas considered in this paper leading to far more sophisticated incentive frameworks [JF06, JF07, Jur07a, Jur07b, JF08] which are most important to improve the above presented mechanism.

The problems which still arise in such reputation mechanisms are manifold and cannot be solved by the reputation mechanism alone. The question arises whether liars should be punished or not. Since detection of false feedback is not always accurate, the system might sometimes punish even truthful agents e.g. if a trustor experienced a defection but the trustee has never defect before and does not defect in the consecutive round. The system will identify the truthful feedback most likely as untrustworthy and punish the “liar” even if what is untrue. Hence, the system would discourage giving feedback because there is a small probability that even truthful reporting is punished. This case is especially relevant if the reporting agent is not payed but rated with a trust rating as in 3.1 and 3.3. Another problem in such reputation mechanisms is connected to the identity of the participants. In an anonymous system we can never be sure that a participant with a very bad reputation, who exploited the system by defecting, starts over by re-entering the system with a “fresh identity”. Another problem that cannot be addressed by the reputation mechanism itself but must be solved by other institutions is collusion.

Agents could try to achieve a better reputation value by making minimal transactions and rate each other positively in order to establish a high reputation they can exploit in the following interactions. One could imagine to weigh the feedback according to the amount of money transferred within the transaction. Still, collusion can take place and has to be inhibited by independent institutions. Therefore, institutions have to be created that punish those kinds of behavior. This should not be an issue that has to be dealt with while setting up the reputation mechanism because it applies not only to one special market or system but to all transactions which are executed in anonymous markets such as the internet.

Table 4: Summary of Reputation Mechanisms

	Liu and Issarny: An Incentive compatible Reputation Mechanism for Ubiquitous Computing Environments	**Jøsang and Ismail: The Beta Reputation System**	**Buchegger and Boudec: A Robust System for P2P and Mobile Ad-hoc Networks**	**Yu and Singh: A Social Mechanism of Reputation Management in electronic Communities**	**Jurca and Faltings: Towards Incentive Compatible Reputation Management**
Ratings	Three different kinds: RRep, SRep and ORep	Reputation rating r_t^x (from X about T)	Two kinds: Reputation rating $R_{i,j}$ and Trust rating $T_{i,j}$ (from i about j)	Trust rating $T_i(j)^t$ (from i about j at time t)	Reputation rating
Elicitation of honest feedback	Judging feedback upon trust rating of the provider and estimating the probability of such behavior with the beta reputation.	Considering the opinion about the provider of information in order to discount the feedback accordingly.	Deviation test checks if the feedback is considered honest.	Different ways of incorporating feedback due to opinions about the provider of information and former transactions with the trustee.	R-Agents check the feedback with the behavior of the concerning agent in the following round.
Incentives	Rating the agents and establishing five states of recommenders; information is shared according to those with different probabilities favoring active, honest recommenders –> incentives through meta-reputation.	No clear incentives.	Incentives through meta reputation ratings but not fully implemented (as done by Liu and Issarny).	No clear incentives.	Payments if report is considered honest.

References

[AG07] Donovan Artz and Yolanda Gil. A survey of trust in computer science and the semantic web. *Web Semant.*, 5(2):58–71, 2007.

[Bü00] Joachim Büschken. *Reputation Networks and "Loose Linkages" between Repuation and Quality.* 2000. Published: Diskussionsbeitrag der Katholischen Universität Eichstätt Wirtschaftswissenschaftliche Fakultät Ingolstadt.

[BB04] Sonja Buchegger and Jean Le Boudec. A robust reputation system for P2P and mobile ad-hoc networks. In *In Proceedings of the Second Workshop on the Economics of Peer-to-Peer Systems*, 2004.

[BKE09] Tina Balke, Stefan König, and Torsten Eymann. A survey on reputation systems for artifcial societies. Bayreuther Arbeitspapiere zur Wirtschaftsinformatik 46, University of Bayreuth, October 2009.

[CP02] Rosaria Conte and Mario Paolucci. *Reputation in Artificial Societies: Social Beliefs for Social Order.* Springer, 1 edition, 2002.

[Del05] Chrysantohos Dellarocas. Reputation mechanisms. 2005.

[JF03] Radu Jurca and Boi Faltings. An incentive compatible reputation mechanism. *In Proceedings of the IEEE Conference on E-Commerce*, page 285—292, 2003.

[JF06] Radu Jurca and Boi Faltings. Minimum payments that reward honest reputation feedback. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 190–199. ACM, 2006.

[JF07] Radu Jurca and Boi Faltings. *Robust Incentive-Compatible Feedback Payments*, pages 204–218. 2007.

[JF08] Radu Jurca and Boi Faltings. Incentives for expressing opinions in online polls. In *Procedings of the 9th ACM Conference on Electronic Commerce*, pages 119–128, Chicago, Il, USA, 2008. ACM.

[JI02] Audun Jøsang and Roslan Ismail. The beta reputation system. *In Proceedings of the 15th Bled Electronic Commerce Conference*, 2002.

[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.

[Jur07a] Radu Jurca. Obtaining reliable feedback for sanctioning reputation mechanisms. *Journal of Artificial Intelligence Research (JAIR)*, 29:391–419, 2007.

[Jur07b] Radu Jurca. *Truthful Reputation Mechanisms for Online Systems*. PhD, Ecole Polytechnique Federale de Lausanne, 2007.

[KHPE08] Stefan König, Sebastian Hudert, Mario Paolucci, and Torsten Eymann. Towards reputation enhanced electronic negotiations for service oriented computing. In Rino et al. Falcone, editor, *TRUST 2008*, volume 5396, pages 273–291, 2008.

[LI06] Jinshan Liu and Valérie Issarny. An incentive compatible reputation mechanism for ubiquitous computing environments. *International Journal of Information Security*, 6(5):297–311, 2006.

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

[RKZF00] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. Reputation systems. *Commun. ACM*, 43(12):45–48, 2000.

[SS05] Jordi Sabater and Carles Sierra. Review on computational trust and reputation models. *Artif. Intell. Rev.*, 24(1):33–60, 2005.

[YS00] Bin Yu and Munindar Singh. A social mechanism of reputation management in electronic communities. In *Cooperative Information Agents IV - The Future of Information Agents in Cyberspace*, pages 154—165. Springer, 2000.

The emergence of the Internet leads to a vast increase in the number of interactions between parties that are completely alien to each other. In general, such transactions are likely to be subject to fraud and cheating. If such systems use computerized rational agents to negotiate and execute transactions, mechanisms that lead to favorable outcomes for all parties instead of giving rise to defective behavior are necessary to make the system work: trust and reputation mechanisms.

This paper examines different incentive mechanisms helping these trust and reputation mechanisms in eliciting users to report own experiences honestly.

ISSN 1864-9300