D2.4.9 Reputation Mechanisms

Radu Jurca (EPFL), Boi Faltings (EPFL),
Walter Binder (EPFL)

Abstract.
EU-IST Network of Excellence (NoE) IST-2004-507482 KWEB
Deliverable D2.4.9 (WP2.4)

This report summarizes the state-of-the-art regarding reputation mechanisms and outlines the reputation model that will be implemented as a prototype in D2.4.6.2. A second version of this document, including a detailed specification of the reputation model, is due in December 2005.

Keyword list: Reputation Mechanisms, Web Services, Semantic Web Services, Semantic Web

Document Identifier	KWEB/2005/D2.4.9/v1.2
Project	KWEB EU-IST-2004-507482
Version	v1.2
Date	August 3, 2005
State	final
Distribution	public

Knowledge Web Consortium

This document is part of a research project funded by the IST Programme of the Commission of the European Communities as project number IST-2004-507482.

University of Innsbruck (UIBK) - Coordinator
Institute of Computer Science
Technikerstrasse 13
A-6020 Innsbruck
Austria
Contact person: Dieter Fensel
E-mail address: dieter.fensel@uibk.ac.at

École Polytechnique Fédérale de Lausanne (EPFL)
Computer Science Department
Swiss Federal Institute of Technology
IN (Ecublens), CH-1015 Lausanne
Switzerland
Contact person: Boi Faltings
E-mail address: boi.faltings@epfl.ch

France Telecom (FT)
4 Rue du Clos Courtel
35512 Cesson Sévigné
France. PO Box 91226
Contact person : Alain Leger
E-mail address: alain.leger@rd.francetelecom.com

Freie Universität Berlin (FU Berlin)
Takustrasse 9
14195 Berlin
Germany
Contact person: Robert Tolksdorf
E-mail address: tolk@inf.fu-berlin.de

Free University of Bozen-Bolzano (FUB)
Piazza Domenicani 3
39100 Bolzano
Italy
Contact person: Enrico Franconi
E-mail address: franconi@inf.unibz.it

Institut National de Recherche en Informatique et en Automatique (INRIA)
ZIRST - 655 avenue de l'Europe -
Montbonnot Saint Martin
38334 Saint-Ismier
France
Contact person: Jérôme Euzenat
E-mail address: Jerome.Euzenat@inrialpes.fr

Centre for Research and Technology Hellas / Informatics and Telematics Institute (ITI-CERTH)
1st km Thermi - Panorama road
57001 Thermi-Thessaloniki
Greece. Po Box 361
Contact person: Michael G. Strintzis
E-mail address: strintzi@iti.gr

Learning Lab Lower Saxony (L3S)
Expo Plaza 1
30539 Hannover
Germany
Contact person: Wolfgang Nejdl
E-mail address: nejdl@learninglab.de

National University of Ireland Galway (NUIG)
National University of Ireland
Science and Technology Building
University Road
Galway
Ireland
Contact person: Christoph Bussler
E-mail address: chris.bussler@deri.ie

The Open University (OU)
Knowledge Media Institute
The Open University
Milton Keynes, MK7 6AA
United Kingdom
Contact person: Enrico Motta
E-mail address: e.motta@open.ac.uk

Universidad Politécnica de Madrid (UPM)
Campus de Montegancedo sn
28660 Boadilla del Monte
Spain
Contact person: Asunción Gómez Pérez
E-mail address: asun@fi.upm.es

University of Karlsruhe (UKARL)
Institut für Angewandte Informatik und Formale Beschreibungsverfahren - AIFB
Universität Karlsruhe
D-76128 Karlsruhe
Germany
Contact person: Rudi Studer
E-mail address: studer@aifb.uni-karlsruhe.de

University of Liverpool (UniLiv)
Chadwick Building, Peach Street
L697ZF Liverpool
United Kingdom
Contact person: Michael Wooldridge
E-mail address: M.J.Wooldridge@csc.liv.ac.uk

University of Manchester (UoM)
Room 2.32. Kilburn Building, Department of Computer Science, University of Manchester, Oxford Road
Manchester, M13 9PL
United Kingdom
Contact person: Carole Goble
E-mail address: carole@cs.man.ac.uk

University of Sheffield (USFD)
Regent Court, 211 Portobello street
S14DP Sheffield
United Kingdom
Contact person: Hamish Cunningham
E-mail address: hamish@dcs.shef.ac.uk

University of Trento (UniTn)
Via Sommarive 14
38050 Trento
Italy
Contact person: Fausto Giunchiglia
E-mail address: fausto@dit.unitn.it

Vrije Universiteit Amsterdam (VUA)
De Boelelaan 1081a
1081HV. Amsterdam
The Netherlands
Contact person: Frank van Harmelen
E-mail address: Frank.van.Harmelen@cs.vu.nl

Vrije Universiteit Brussel (VUB)
Pleinlaan 2, Building G10
1050 Brussels
Belgium
Contact person: Robert Meersman
E-mail address: robert.meersman@vub.ac.be

Work package participants

The following partners have taken an active part in the work leading to the elaboration of this document:

Ecole Polytechnique Fédérale de Lausanne

Changes

Version	Date	Author	Changes
0.5	01.03.05	Walter Binder	creation
0.6	15.03.05	Walter Binder	state-of-the-art added
0.7	12.04.05	Walter Binder	state-of-the-art extended
0.8	20.05.05	Walter Binder	model overview and scenarios added
1.0	21.06.05	Walter Binder	finalization of first version
1.1	31.07.05	Walter Binder	addressed review comments
1.2	03.08.05	Radu Jurca	addressed further review comments

Executive Summary

In an open environment where malicious parties may advertise false service capabilities the use of reputation services is a promising approach to mitigate such attacks. Misbehaving services receive a bad reputation and will be avoided by other clients. Reputation mechanisms help to improve the global efficiency of the overall system because they reduce the incentive to cheat.

This report summarizes the state-of-the-art regarding reputation mechanisms. We consider ways of modeling trust, computational models of trust, as well as incentive-compatible reputation mechanisms. Concerning computational models of trust, we distinguish social trust networks, probabilistic estimation techniques, and game-theoretic models. For these approaches, we consider different feedback aggregation strategies.

We also outline a reputation model that will be further developed in the second version of this document (due in December 2005) and implemented as a prototype in D2.4.6.2.

Contents

1 Introduction **1**

2 State of the Art **3**
2.1 Modeling Human Trust 4
2.2 Computational Models of Trust 5
2.2.1 Social Trust Networks 5
2.2.2 Probabilistic Estimation Techniques 6
2.2.3 Game-Theoretic Models 7
2.2.4 Feedback Aggregation Strategies 7
2.3 Incentive Compatibility 9

3 Selected Reputation Model **11**
3.1 Overview of Reputation Model 12
3.2 Example Scenarios 15
3.3 Modeling and Implementation Techniques 16

4 Conclusion **17**

Chapter 1

Introduction

The availability of ubiquitous communication through the Internet is driving the migration of commerce and business from direct interactions between people to electronically mediated interactions. It is enabling a transition to peer-to-peer commerce without intermediaries and central institutions.

Most business transactions have the form of prisoners' dilemma games where dishonest behaviour is the optimal strategy. For instance, in a service level agreement, there is no incentive for the service provider to deliver the promised quality of service once he has received the client's payment (assuming the absence of a public security infrastructure that mediates every transaction). In conventional commerce, personal relations create psychological barriers against such behaviour. In electronically mediated peer-to-peer commerce, there is no physical contact and even identities can be easily faked. Fraud and deception are a major obstacle to realizing the huge economic benefits of peer-to-peer commerce.

A standard approach in traditional business to avoid cheating (i.e., to avoid deviation from a promise / from partners' expectations) is to use trusted third parties (TTP) that oversee the transactions and rule out or at least punish cheating. In electronic interactions this approach is not always possible, as they pose problems of verification, scalability, cost and legality when several countries are involved in transactions.

In this report, we pursue a fundamentally different approach to electronically mediated business that can do without enforcement by third parties. We consider reporting, sharing and using reputation information in a network of agents as part of a mechanism that makes cooperation the dominant strategy in business transactions. Such an approach aims at re-establishing a social framework that supports trusted interactions. It is based on the observation that agent strategies change when we consider that interactions are repeated: the other party will remember past cheating, and change its terms of business accordingly in the future. In this case, the expected future gains due to future transactions can offset the loss incurred by not cheating in the present transaction [39]. This effect can be amplified considerably if such reputation information is shared among a large population and thus

multiplies the expected future gains made accessible by honest behaviour. Game theorists have studied the reputation effect for many years and established results that show its feasibility in a wide variety of scenarios. What is missing now are robust and scalable computational mechanisms for implementing them in electronic peer-to-peer commerce.

This report is structured as follows: In Chapter 2 we discuss the current state-of-the-art concerning reputation mechanisms. We address issues regarding the modeling of trust as perceived by human beings, computational models of trust, and incentive-compatible reputation mechanisms. In Chapter 3 we sketch the reputation model that we will further develop in the second version of this document and implement in D2.4.6.2. Finally, Chapter 4 concludes this report.

Chapter 2

State of the Art

We see trust related research as going into three directions:

1. Work that models the notion of "real world" trust (as used in sociology and psychology primarily) and propose definitions of trust that are appropriate for use in online settings. The definition of trust and its corresponding meaning is a much disputed issue among the computer science community. Since the human understanding of the notion of trust is much too complex to be modelled within an artificial system, authors usually consider just facets of the notion of trust, and define it corresponding to their needs. Trust modeling is addressed in Section 2.1.

2. Computational models of trust, proposing concrete models for trust evaluation. We characterize these models along the following two dimensions:

   - How precisely they boost trust in the community in which they are deployed.
   - How efficiently they can be implemented in a decentralized network of agents.

   Computational models of trust are discussed in Section 2.2.

3. Incentive compatibility related works. Incentive compatibility is one of the most desirable properties of protocols involving communication among autonomous, self-interested agents. Its existence assures that specific behaviour (truth telling, in particular) is equilibrium of the game constructed from the protocol. Applied to the trust models, incentive compatibility would imply truthful reporting of reputation information. Incentive-compatible reputation mechanisms are covered by Section 2.3.

2.1 Modeling Human Trust

Generally, the notion of trust is used to refer to a subjective decision making process that takes into consideration a lot of factors. [30] explains how human beings deal with the trust decision making process by using a set of rules. The model proposed is the Social Auditor Model. The author also studies the efficiency of different rules and strategies that can be used within an artificial society.

In [5, 21, 14] the authors look at the dynamics associated with the notion of trust. Trust and distrust responsiveness (trust from the trustee increases the probability of cooperative behaviour from the trustor, while distrust from the trustee increases the probability of defective behaviour) are presented as facts of human behaviour. Also the dialectic link between trust and degree of control is addressed.

In [33] a multi-disciplinary literature survey on the notion of trust and distrust is presented. The paper develops a conceptual topology of the factors that contribute towards trust and distrust decisions and defines as subsets of the high level concepts measurable constructs for empirical research.

One of the input information that is often used in a trust decision making process is the reputation of the partner. Reputation can be regarded as a unitary appreciation of the personal attributes of the trustor: competence, benevolence, integrity and predictability. [36] presents an extensive classification of reputation by the means of collecting it. Experiments for finding out which component contributes the most towards correct trust decisions are also conducted.

As belonging to this group can be regarded works that investigate some inherent characteristics of the online world that any trust management model must be aware of. [22] discusses risks associated with the ease at which members of online communities can change their identities. Through a game theoretic modelling, they come to a conclusion that newcomers must start with the lowest possible reputation value in order to be discouraged to misbehave and change their identity afterwards. [15] identifies a number of possible attacks on reputation reporting systems (“ballot stuffing”, “bad mouthing”, etc.) and proposes an appropriate solution to reduce effects of those attacks. [40] identifies main patterns of human behaviour with respect to trust. It argues that, despite clear incentives to free ride (not leave feedback) and leave only positive feedback, trust among eBay traders emerges due to its reputation system.

While humans address trust issues in a complex way, considering the direct experience with a provider, the experience of others with the provider, as well as social aspects (e.g., nationality, group membership, etc.), only one aspect – the reputation of the provider – is modeled in computer systems, as discussed in the following section.

2.2 Computational Models of Trust

In the categorization of the computational models of trust we will adopt here a division based on how they perform the goal of bootstrapping trust. We see the following three broad classes of approaches: 1) social (trust) networks formation (Section 2.2.1), 2) probabilistic estimation techniques (Section 2.2.2), and 3) game-theoretic models (Section 2.2.3). For all these approaches, different feedback aggregation strategies are possible, as discussed in Section 2.2.4.

2.2.1 Social Trust Networks

The underlying assumption of the class of social networks formation is that the agents engage in bilateral interactions whose outcomes are evaluated and aggregated, which results in forming a trust graph in which each branch (a,b) is assigned a weight representing the trust of agent a towards agent b aggregated across all interactions between them in which agent a happened to have relied on agent b. Having the local interactions among the agents encoded this way, the challenge is how to merge these local beliefs to enable the agents to compute the trustworthiness of non-neighbouring agents, whom they never met before. The main distinguishing points among the numerous works belonging to this class are: 1) the strategy to aggregate individual experiences to give the mentioned weights, 2) the strategy to aggregate the weights along a path of an arbitrary length to give a path wide external opinion and 3) the strategy to aggregate this external opinion across multiple paths between two given agents.

[7] presents an early example in which a clear distinction between direct experiences and recommendations has been made, which is reflected in the strategy for path wide external opinion aggregation. However, this separation of the two contexts led to an exponential complexity of the trust derivation algorithm. Clearly, this is unacceptable for large scale networks.

[50] does not treat recommendations and direct service provisions separately. It uses a variation of the delta learning method to aggregate "positive" and "negative" experiences of the agents into the weights assigned to the corresponding branches and simple multiplication as the strategy to compute the path wide external opinions. As for the strategy to aggregate the external opinion of different paths the authors use a variation of the simple maximum function. All this results in a polynomial time algorithm for the overall trust aggregation.

[41] offers important theoretical insights on how the computational complexity of the trust derivation algorithms relates to the mentioned aggregation strategies by characterizing the combinations of path and across-path aggregation strategies that may lead to a non-exponential trust computation algorithm (we note that many other works use such combinations: e.g., [38] and [29]). The authors also offer such an algorithm which is, however, based on a synchronous participation of all agents in the network. As such it

is not quite appropriate for usage in P2P networks due to their inherent high dynamicity. With respect to this problem [49] offers a considerable improvement in terms of an appropriate caching scheme that enables asynchronous computation while retaining good performance. A common denominator of all these works is that the computed values have unclear semantics and are hard to interpret on an absolute scale, without ranking them. In many applications this imposes certain problems. On the other hand, as shown by many simulations, they are very robust to a wide range of misbehaviours.

2.2.2 Probabilistic Estimation Techniques

Probabilistic estimation techniques present certain improvement with respect to the meaningfulness of the computed values. Namely, they output probability distributions (or at least the most likely outcome) over the set of possible behaviours of the trusted agents enabling thus the trusting agents to evaluate explicitly their utilities from the decision to trust or not. [37] presents the well-known method of Bayesian estimation as the right probabilistic tool for assessing the future trusting performance based on past interactions. Only direct interactions were studied - the question of including recommendations was not considered. [13] goes a step further by taking into account the "second-hand" opinions also. However, the strategy for merging own experiences with those of other witnesses is intuitive (giving more weight to own experiences, though plausible, is still intuitive) rather than theoretically founded.

[3] presents a decentralized trust management model that analyzes past interactions among agents to make a probabilistic assessment of whether any given agent cheated in his past interactions. The emphasis is put not only on assessing trust but also on providing a scalable data management solution particularly suitable for decentralized networks. The problem in decentralized networks is that the reputation data is aggregated along wrong dimension in the sense that each agent has information about his own past interactions with others but cannot easily obtain opinions of others about any other particular agent in the network. To achieve the needed reaggregation of reputation data, the authors use P-Grid, a scalable data access structure for P2P networks [2]. For any particular agent, they designate a set of replicas to store the feedbacks, ratings of trusting behaviour of that agent (complaints filed by him about others and complaints filed by others about him) so that the reputation data can be accessed and collected efficiently, in logarithmic time. As replicas may provide false data, an appropriate replication factor along with a proper voting scheme to choose the most likely reputation data set are chosen in order to achieve accurate predictions. Trust assessments themselves are made based on an analysis of agent interactions modelled as Poisson processes. As was shown by simulations, cheating behaviour of the agents can be identified with a very high probability. The model is simplistic in the sense that, for any agent, it outputs whether the agent cheated in the past or not, but it can be easily extended to give predictions of the agents' trusting behaviour, as done e.g. in [49].

[19] is a step further towards analyzing how trust predictions can be used in the context of making business decisions. Safe exchange represents an approach to gradual exchanges of goods and money in which both payments and goods are chunked with their deliveries scheduled in such a way that both exchange partners are better off by continuing the exchange till its end than by breaking it at any step before. The authors provide a trust aware extension of the original approach [42] by modelling trust explicitly.

[20] proposes a double auctioning mechanism that does not rely on the existence of central authorities, auctioneer in particular. As such it is amenable to implementation in P2P environments. The mechanism has good economic properties such as, for example, fast convergence towards efficient trading through intuitive and simple bidding strategies. However, these properties can only be guaranteed in the presence of a distributed reputation mechanism.

2.2.3 Game-Theoretic Models

Game-theoretic reputation models make a further clarification in the interpretation of the agents' trustworthiness in the sense that, if the reputation system is designed properly, trust is encoded in the equilibria of the repeated game the agents are playing. Thus, for rational players trustworthy behaviour is enforced. The real challenge here is how to define the feedback aggregation strategies that will lead to socially desirable outcomes carrying trust.

Theoretic research on reputation mechanisms started with the seminal papers of Kreps, Milgrom, Wilson and Roberts [31, 32, 34] who explained how a small amount of incomplete information is enough to generate the reputation effect, (i.e., the preference of agents to develop a reputation for a certain type) in the finitely repeated Prisoners' Dilemma game and Selten's Chain-Store game [45].

Fudenberg and Levine [23] and Schmidt [44] continue on the same idea by deriving lower bounds on the equilibrium payoff received by the reputable agent in two classes of games in which the reputation effect can occur.

[18] focuses on a specific game and derives its equilibria. Apart from this the author also raises questions concerning the overall game-theoretic reputation systems design, such as incentivizing players to leave feedback, dealing with incomplete feedback etc. However, an underlying assumption of this work is that a central trusted authority does the feedback aggregation. We see this as a major obstacle to transferring game-theoretic models to decentralized environments.

2.2.4 Feedback Aggregation Strategies

For the previously mentioned approaches, there are different strategies to aggregate the external opinion. [15, 16, 53, 1] use collaborative filtering techniques to calculate person-

alized reputation estimates of as weighted averages of past ratings in which weights are proportional to the similarity between the agent who computes the estimate and the raters.

[8, 9, 10] describe computational trust mechanisms based on direct interaction-derived reputation. Agents learn to trust their partners, which increases the global efficiency of the market. However, the time needed to build the reputation information prohibits the use of this kind of mechanisms in a large scale online market.

[48] uses machine learning techniques and heuristic methods to increase the global performance of the system by recognizing and isolating defective agents. Common to these works is that they consider only direct reputation. Similar techniques, extended to take into account indirect reputation, are used in [6, 43, 26, 51, 46].

A number of reputation mechanisms also take into consideration indirect reputation information, i.e., information reported by peers. [43, 52] use social networks in order to obtain the reputation of an unknown agent. Agents ask their friends, who in turn can ask their friends about the trustworthiness of an unknown agent. Recommendations are afterwards aggregated into a single measure of the agent's reputation. This class of mechanisms, however intuitive, does not provide any rational participation incentives for the agents. Moreover, there is little protection against untruthful reporting, and no guarantee that the mechanism cannot be manipulated by a malicious provider in order to obtain higher payoffs.

In [26] the authors present an example of a reputation sharing mechanism that is also incentive-compatible (see Section 2.3 for details). The mechanism is based on side payments that are organized through a set of broker agents called R-agents, which buy and sell reputation information. A simple payment rule (incoming reputation reports are paid only if they mach the next reputation report filed about the same agent) makes it rational for agents to truthfully share reputation information. The mechanism is decentralized and robust (up to certain limits) to irrational untruthful reporting.

2.3 Incentive Compatibility

The vast majority of the previously discussed models are not fully incentive-compatible in the sense that it is in the best interest of all agents to report their feedbacks truthfully (and leave feedback whatsoever, without an incentive to free ride). The closest to the idea of incentive-compatibility are [11, 12, 35, 25, 28].

[11] considers exchanges of goods for money and proves that markets in which agents are trusted to the degree they deserve to be trusted is equally efficient as a market with complete trustworthiness. It then presents an exchange scenario in which buyers announce their trustworthiness and sellers compute their estimates of the same. It was shown that, with an appropriately chosen advance payment, buyers cannot benefit from announcing false levels of trustworthiness. We must note that the generalness of the results is somewhat limited by the assumption that the contract price is chosen according to a particular bargaining solution (Nash's bargaining solution in this case). For auctions which are not completely enforceable, the same authors describe in [12] a mechanism based on discriminatory bidding rules that separate trustworthy from untrustworthy bidders. Again, this result has the same limitations as mentioned above.

For e-Bay-like auctions, the Goodwill Hunting mechanism [17] provides a way in which the sellers can be made indifferent between lying or truthfully declaring the quality of the good offered for sale. Momentary gains or losses obtained from misrepresenting the good's quality are later compensated by the mechanism which has the power to modify the announcement of the seller.

A significant contribution towards eliciting honest reporting behaviour is made in [35]. The authors propose scoring rules as payment functions which induce rational honest reporting. The scoring rules however, cannot be implemented without accurately knowing the parameters of the agents' behaviour model, which can be a problem in real-world systems. Moreover, this mechanism works only when the set of possible seller types is countable and contains at least 2 elements, and when the signals received by the buyers about the seller's behaviour are independently identically distributed from one interaction to the other.

Using the same principle, [25] overcomes the need to know the parameters of the agents' behaviour model at the expense of further reducing the acceptable provider behaviour types. [28] describes a novel protocol to elicit truthful reputation information in electronic markets lacking independent verification authorities by correlating the reports of the seller and buyer involved in the same transaction.

[25] further strengthens this result by theoretically delimiting the minimum set of conditions necessary for the mechanism to be incentive-compatible. Moreover, using digital signatures and a contracting protocol, the authors show how the mechanism can be made secure against identity theft (agents stealing the identity of other agents in order to benefit from an undeserved reputation) and manipulation by any single agent. A concrete implementation of this mechanism is deployed on the Agentcities platform [4].

As opposed to side-payment schemes that correlate a present report with future reports submitted about the same agent, [28] presents a mechanism that discovers (in equilibrium) the true outcome of a transaction by analyzing the two reports coming from the agents involved in the exchange. For two long-run rational agents, the authors show that it is possible to design such a mechanism that makes cooperation a stable equilibrium. The mechanism involves no independent verification authority, and is easily distributable as the decision about the true outcome of a transaction does not depend on any past or future interactions.

Chapter 3

Selected Reputation Model

In this chapter we outline the reputation model that we selected for implementation in D2.4.6.2. In this report, we focus only on the requirements concerning the reputation model. The details of our reputation model will be presented in the second version of this document (due in December 2005) and a prototype implementation will be available as part of D2.4.6.2.

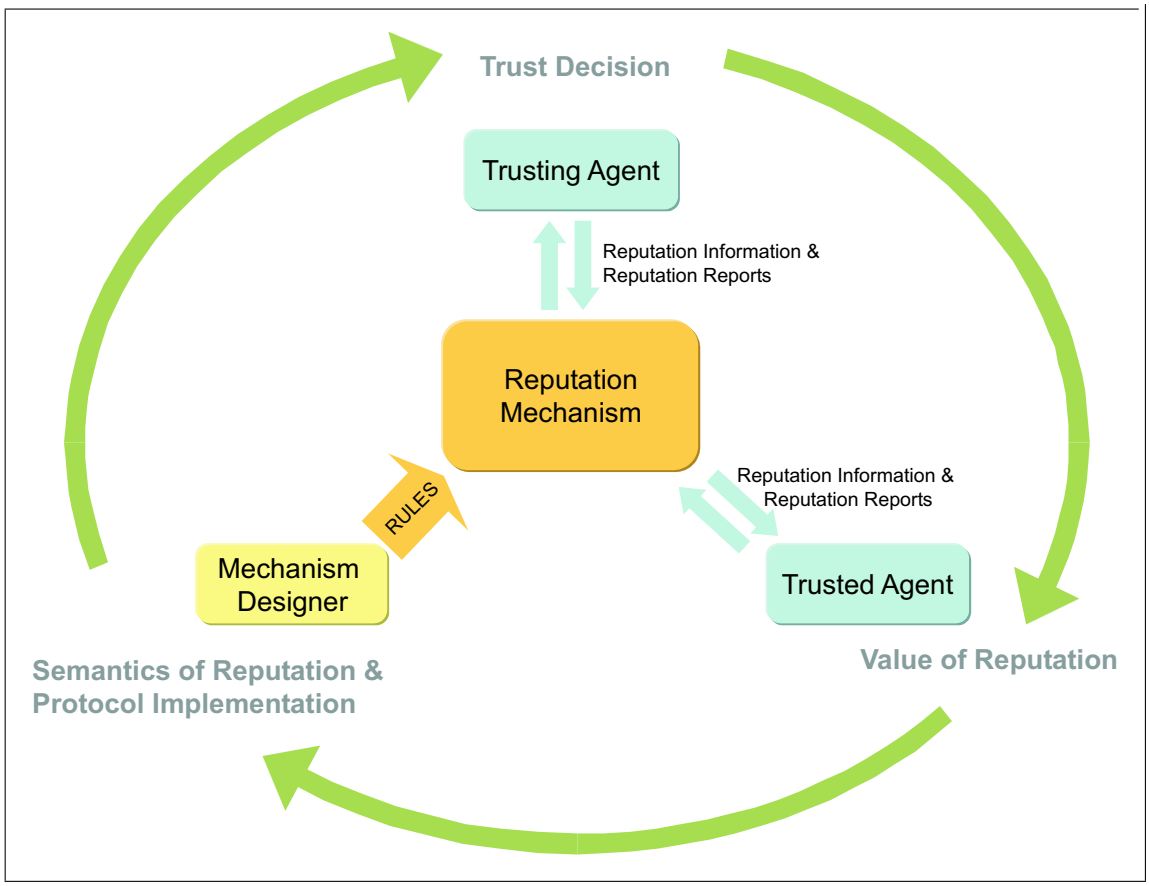

Figure 3.1: This figure illustrates the reputation mechanism as an equilibrium solution to interrelated and conflicting aspects. Note that the reputation mechanism could be physically implemented by the same agent using the mechanism.

3.1 Overview of Reputation Model

We consider the setting shown in Figure 3.1, where trust between trusting and trusted agents is mediated by a reputation mechanism designed by a mechanism designer. The reputation mechanism appears as an equilibrium solution to the following three interrelated questions:

1. How do trusting agents use the reputation information in order to make trust decisions regarding trusted agents?
2. What value do rational trusted agents associate with reputation information? This value depends on the trusting decision of the trusting agent and reflects the influence reputation has on future gains.
3. How can a designer build a reputation mechanism such that the value of reputation for trusted agents outweighs the momentary gain obtained when cheating?

None of these questions can be answered separately. They need to be treated together such that a stable solution is reached (i.e., the answer to one question does not trigger a change in the solution of the next).

We concentrate on the two main aspects of a reputation mechanism: (1) the semantics of reputation and (2) the protocol implementation. A clear semantics of reputation is given by deciding on:

- The type of feedback collected by the mechanism: i.e., what kind of information about an agent's past behaviour is relevant for a trusting agent in making trust decisions. The type of feedback is context dependent.
- Feedback aggregation rules: i.e., how can feedback be aggregated into meaningful reputation information.

Well defined reputation information results in clear guidelines for taking trust decisions based on reputation and for evaluating the value of reputation.

Regarding protocol implementation, a number of functional properties have to be taken into account, that determine the quality and the reliability of the mechanism proposed:

1. Incentive-compatibility: agents should have the incentive to provide truthful feedback to the mechanism.
2. Security against non-collusion: an important aspect of multi-agent systems is that agents are independent and do not collude. Certain protocols are robust against collusion, while for others collusion has to be ruled out for example by randomisation or cryptographic mechanisms.
3. Scalability: implementing the mechanism in a large network of independent, self-organizing agents leads to non-trivial scalability problems in terms of communication and information management cost.
4. Robustness: in real-world settings robustness of the mechanism against failure, faulty information and malicious behaviour cannot be ignored.
5. Bounded rationality: the limits of knowledge and computation time influence agent strategies and hence the entire mechanism.

The main research methodology will be to develop and implement methods that address in particular issues of incentive-compatibility and security against collusion. For experimental evaluation of the methods we will develop working prototypes and validate them. An iterative development process will be employed in which the experimental results will be used to refine the methods and start a new iteration until a stable point is

reached in which no further refinement is needed. Our goal is not only to gain an understanding of the possibilities and issues regarding reputation mechanisms in distributed environments, but also to propose algorithms and evaluate them on implementations and with human participants.

3.2 Example Scenarios

Due to the complexity of the interactions in distributed environments, we do not expect to be able to evaluate all properties analytically. Therefore, we will develop a decentralized software platform and several example scenarios in which reputation mechanisms can be empirically evaluated and compared through simulation. We will develop demonstration scenarios and for each of them a software implementation with facilities for gathering statistics that allows simulating different reputation mechanisms and agent strategies. The scenarios will be selected so that they illustrate the different kinds of environments that are being constructed in the Internet, such as for example:

- A P2P file sharing service: Agents need to trust the quality and the availability of the shared resource, media files in this case. The risk lies in not being satisfied with the downloaded file, loose more time and energy to find another provider, spend some more on downloading costs. Reputation information is based on the quality of service and availability. No exchange of money occurs.
- A decentralized electronic auction marketplace: The buyer agent needs to trust that the seller will correctly describe the product offered for sale and that he will cooperate and actually ship the product. Reputation information is based on honesty (correctly reporting the attributes of the product), predictability (degree to which the agent will behave as predicted), and integrity (agent makes deals in good faith). On the other hand, the reputation mechanism itself must be trustworthy in order to ensure the proper functioning of the market.
- A marketplace for web services: Consumers need to trust that service providers will offer the promised service at the promised quality level. Reputation information is context-dependent, i.e., depends on the type of service requested.
- A marketplace for financial information: Decision makers need to trust that the information on which they base their decisions is correct. A particularity of this application is that conflicts of interests among organizations may strategically influence the quality of provided information.

We will first analyze the scenarios and the interaction patterns among the agents to determine what role trust plays in them. For example, trust can be about the quality of information, about the reliability of a service, about the honesty of an agent, about the predictability of its behaviour, etc. Then, we will consider what kind of reputation information can be used to improve such trust estimates as well as how relevant information can be gathered by the agents. The output of this task will provide a full characterization of the above scenarios along the following four dimensions: (i) interaction type, (ii) value of trust, (iii) relevant reputation information, and (iv) strategies to collect reputation information.

3.3 Modeling and Implementation Techniques

Simple, binary feedback mechanisms can be deployed in many environments. They are intuitive, easy to use, and require the least effort from the reporter. We will develop theoretical models for aggregating binary feedback into semantically well defined reputation information.

We will proceed by decomposing the problem in two distinct steps. First, we will study possible models of binary reputation mechanisms when reporting agents are cooperative (i.e. they do not explicitly try to manipulate the reputation mechanism), however not entirely reliable (i.e. they can make unintentional mistakes when submitting feedback). For this step, the research methodology is the following:

1. Assume a setting in which reporters always tell the truth and undistorted information is available to the reputation mechanism.
2. Derive feedback aggregation and trust decision rules that assign a value to a feedback report such that the momentary gain obtained from cheating is offset by the loss due to negative feedback.
3. Study the effect of mistakes and imperfect information on the value of a reputation report.
4. Validate the theoretical results in 2 and 3 against a simulated environment in which reporters make unintentional mistakes.

Second, we will investigate the resulting models when agents strategically try to manipulate the mechanism. The objective of this second step is to enhance reputation mechanisms with interaction protocols that make them incentive-compatible (i.e. rational agents have the incentive to report the truth) and secure against manipulation by other agents.

Due to computational limitations or to the inherent uncertainty of the environment, the behavior of many service providers (trusted agents) can be modelled by Markov Chains [47]. For such behavior models we will develop interaction protocols based on side-payments and cryptographic methods that (1) make it in the best interest of the reporting agents to file true feedback and (2) make it impossible (i.e., sufficiently expensive) for agents to manipulate the reputation of any single provider. The methodology used for achieving this goal involves iteratively:

- Designing interaction protocols achieving the desired properties.
- Validating them in simulated distributed environments in which mistakes and failures can occur.

Chapter 4

Conclusion

This report gave an overview of the state-of-the-art concerning reputation mechanisms and outlined the design of a particular incentive-compatible reputation model that will be further specified and implemented in KnowledgeWeb WP 2.4.

In an open environment where malicious parties may advertise false service capabilities the use of reputation services is a promising approach to mitigate such attacks. Misbehaving services receive a bad reputation (reported by disappointed clients) and will be avoided by other clients. Reputation mechanisms help to improve the global efficiency of the overall system because they reduce the incentive to cheat [9]. Studies show that buyers seriously take into account the reputation of the seller when placing their bids in online auctions [24]. Moreover, it has been proven that in certain cases reputation mechanisms can be designed in such a way that it is in every party's interest to report correct reputation information (incentive compatible reputation services) [27]. Besides, reputation mechanisms can be implemented in a secure way [25].

Considering the importance of reputation services in open environments, it is essential that service discovery and composition algorithms intended to operate in such environments exploit these reputation services in order to favor services with a high reputation. In the second version of this document (due in December 2005) we will define the interfaces of our reputation mechanism, which will be provided as a web service. An implementation of our reputation mechanism will be part of the prototype developed in D2.4.6.2.

Bibliography

[1] A. Abdul-Rahman and S. Hailes. Supporting Trust in Virtual Communities. In *Proceedings Hawaii International Conference on System Sciences*, Maui, Hawaii, 2000.

[2] K. Aberer. P-Grid: A self-organizing access structure for P2P information systems. *Lecture Notes in Computer Science*, 2172, 2001.

[3] K. Aberer and Z. Despotovic. Managing Trust in a Peer-2-Peer Information System. In *Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM)*, 2001.

[4] Agentcities. http://www.agentcities.net.

[5] M. Bacharach. How Human Trusters Assess Trustworthiness in Quasi-Virtual Contexts. In *Proceedings of the AAMAS Workshop on Trust Deception and Fraud*, Bologna, Italy, 2002.

[6] S. Barber and J. Kim. Belief Revision Process Based on Trust: Agents Evaluating Reputation of Information Sources. In R. Falcone, M. Singh, and Y.-H. Tan, editors, *Trust in Cyber-societies*, volume LNAI 2246, pages 73–82. Springer-Verlag, Berlin Heidelberg, 2001.

[7] T. Beth, M. Borcherding, and B. Klein. Valuation of Trust in Open Networks. In *Proceedings of the European Symposium on Research in Computer Security (ESORICS)*, pages 3–18, Brighton, UK, 1994. Sprinter-Verlag.

[8] A. Birk. Boosting Cooperation by Evolving Trust. *Applied Artificial Intelligence*, 14:769–784, 2000.

[9] A. Birk. Learning to Trust. In R. Falcone, M. Singh, and Y.-H. Tan, editors, *Trust in Cyber-societies*, volume LNAI 2246, pages 133–144. Springer-Verlag, Berlin Heidelberg, 2001.

[10] A. Biswas, S. Sen, and S. Debnath. Limiting Deception in a Group of Social Agents. *Applied Artificial Intelligence*, 14:785–797, 2000.

[11] S. Braynov and T. Sandholm. Incentive Compatible Mechanism for Trust Revelation. In *Proceedings of the AAMAS*, Bologna, Italy, 2002.

[12] S. Braynov and T. Sandholm. Auctions with Untrustworthy Bidders. In *Proceedings of the IEEE Conference on E-Commerce*, Newport Beach, CA, USA, 2003.

[13] S. Buchegger and J.-I. L. Boudec. The effect of rumour spreading in reputation systems for mobile ad-hoc networks. In *Proceedings of WiOpt '03: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, Sophia-Antipolis, France, Mar. 2003.

[14] C. Castelfranchi and R. Falcone. Trust and Control: A Dialectic Link. *Applied Artificial Intelligence*, 14:799–823, 2000.

[15] C. Dellarocas. Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behaviour. In *Proceedings of the 2nd ACM conference on Electronic Commerce*, Minneapolis, MN, 2000.

[16] C. Dellarocas. The Design of Reliable Trust Management Systems for Electronic Trading Communities. Working Paper, MIT, 2001.

[17] C. Dellarocas. Goodwill Hunting: An Economically Efficient Online Feedback. In J. Padget and et al., editors, *Agent-Mediated Electronic Commerce IV. Designing Mechanisms and Systems*, volume LNCS 2531, pages 238–252. Springer Verlag, 2002.

[18] C. Dellarocas. Efficiency and Robustness of Binary Feedback Mechanisms in Trading Environments with Moral Hazard. MIT Sloan Working Paper #4297-03, 2003.

[19] Z. Despotovic and K. Aberer. Trust-aware delivery of composite goods. In *AP2PC*, pages 57–68, 2002.

[20] Z. Despotovic, J.-C. Usunier, and K. Aberer. Towards peer-to-peer double auctioning. In *HICSS*, 2004.

[21] R. Falcone and C. Castelfranchi. The Socio-cognitive Dynamics of Trust: Does Trust create Trust. In R. Falcone, M. Singh, and Y.-H. Tan, editors, *Trust in Cyber-societies*, volume LNAI 2246, pages 55–72. Springer-Verlag, Berlin Heidelberg, 2001.

[22] E. Friedman and P. Resnick. The Social Cost of Cheap Pseudonyms. *Journal of Economics and Management Strategy*, 10(2):173–199, 2001.

[23] D. Fudenberg and D. Levine. Reputation and Equilibrium Selection in Games with a Patient Player. *Econometrica*, 57:759–778, 1989.

[24] D. Houser and J. Wooders. Reputation in Internet Auctions: Theory and Evidence from eBay. University of Arizona Working Paper #00-01, 2001.

[25] R. Jurca and B. Faltings. An Incentive-Compatible Reputation Mechanism. In *Proceedings of the IEEE Conference on E-Commerce*, Newport Beach, CA, USA, 2003.

[26] R. Jurca and B. Faltings. Towards Incentive-Compatible Reputation Management. In R. Falcone, R. Barber, L. Korba, and M. Singh, editors, *Trust, Reputation and Security: Theories and Practice*, volume LNAI 2631, pages 138 – 147. Springer-Verlag, Berlin Heidelberg, 2003.

[27] R. Jurca and B. Faltings. "CONFESS". An Incentive Compatible Reputation Mechanism for the Online Hotel Booking Industry. In *Proceedings of the IEEE Conference on E-Commerce*, San Diego, CA, USA, 2004.

[28] R. Jurca and B. Faltings. Truthful reputation information in electronic markets without independent verification. Technical Report ID: IC/2004/08, EPFL, http://ic2.epfl.ch/publications, 2004.

[29] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. EigenRep: Reputation management in P2P networks. In *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary, May 2003.

[30] R. Kramer. Trust Rules for Trust Dilemmas: How Decision Makers Think and Act in the Shadow of Doubt. In R. Falcone, M. Singh, and Y.-H. Tan, editors, *Trust in Cyber-societies*, volume LNAI 2246, pages 9–26. Springer-Verlag, Berlin Heidelberg, 2001.

[31] D. M. Kreps, P. Milgrom, J. Roberts, and R. Wilson. Rational Cooperation in the Finitely Repeated Pisoner's Dilemma. *Journal of Economic Theory*, 27:245–252, 1982.

[32] D. M. Kreps and R. Wilson. Reputation and Imperfect Information. *Journal of Economic Theory*, 27:253–279, 1982.

[33] H. McKnight and N. Chervany. Trust and Distrust: One Bite at a Time. In R. Falcone, M. Singh, and Y.-H. Tan, editors, *Trust in Cyber-societies*, volume LNAI 2246, pages 27–54. Springer-Verlag, Berlin Heidelberg, 2001.

[34] P. Milgrom and J. Roberts. Predation, Reputation and Entry Deterrence. *J. Econ. Theory*, 27:280–312, 1982.

[35] N. Miller, P. Resnick, and R. Zeckhauser. Eliciting Honest Feedback in Electronic Markets. Working Paper, 2003.

[36] L. Mui, A. Halberstadt, and M. Mohtashemi. Notions of Reputation in Multi-Agents Systems:A Review. In *Proceedings of the AAMAS*, Bologna, Italy, 2002.

[37] L. Mui, M. Mohtashemi, and A. Halberstadt. A Computational Model of Trust and Reputation. In *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS)*, 2002.

[38] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, Stanford University, Stanford, CA, USA, Nov. 1998.

[39] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. Reputation systems. *Communications of the ACM*, 43(12):45–48, Dec. 2000.

[40] P. Resnick and R. Zeckhauser. Trust Among Strangers in Electronic Transactions: Empirical Analysis of eBay's Reputation System. In M. Baye, editor, *The Economics of the Internet and E-Commerce*, volume 11 of Advances in Applied Microeconomics. Elsevier Science, Amsterdam, 2002.

[41] M. Richardson, P. Domingos, and R. Agrawal. Trust management for the semantic web. In *Proceedings of the Second International Semantic Web Conference*, pages 351–368, Sanibel Island, FL, USA, Sept. 2003.

[42] T. W. Sandholm. *Negotiation among self-interested computationally limited agents*. PhD thesis, University of Massachusetts at Amherst, 1996.

[43] M. Schillo, P. Funk, and M. Rovatsos. Using Trust for Detecting Deceitful Agents in Artificial Societies. *Applied Artificial Intelligence*, 14:825–848, 2000.

[44] K. M. Schmidt. Reputation and Equilibrium Characterization in Repeated Games with Conflicting Interests. *Econometrica*, 61:325–351, 1993.

[45] R. Selten. The Chain-Store Paradox. *Theory and Decision*, 9:127–159, 1978.

[46] S. Sen and N. Sajja. Robustness of Reputation-based Trust: Boolean Case. In *Proceedings of the AAMAS*, Bologna, Italy, 2002.

[47] N. Sokey and R. Lucas. *Recursive Methods in Economic Dynamics*. Harvard University Press, 1989.

[48] M. Witkowski, A. Artikis, and J. Pitt. Experiments in building Experiential Trust in a Society of Objective-Trust Based Agents. In R. Falcone, M. Singh, and Y.-H. Tan, editors, *Trust in Cyber-societies*, volume LNAI 2246, pages 111–132. Springer-Verlag, Berlin Heidelberg, 2001.

[49] L. Xiong and L. Liu. Peertrust: Supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Trans. Knowl. Data Eng.*, 16(7):843–857, 2004.

[50] B. Yu and M. Singh. A Social Mechanism of Reputation Management in Electronic Communities. In *Proceedings of the Forth International Workshop on Cooperative Information Agents*, pages 154–165, 2000.

[51] B. Yu and M. Singh. An Evidential Model of Distributed Reputation Management. In *Proceedings of the AAMAS*, Bologna, Italy, 2002.

[52] B. Yu and M. Singh. Detecting Deception in Reputation Management. In *Proceedings of the AAMAS*, Melbourne, Australia, 2003.

[53] G. Zacharia, A. Moukas, and P. Maes. Collaborative Reputation Mechanisms in Electronic Marketplaces. In *Proceedings of the 32nd Hawaii International Conference on System Sciences (HICSS)*, 1999.