Available online at www.sciencedirect.com
ScienceDirect
Electronic Notes in Theoretical Computer Science
ELSEVIER
Electronic Notes in Theoretical Computer Science 197 (2008) 73–89
www.elsevier.com/locate/entcs

The Case for Fairness of Trust Management

Adam Wierzbicki[1]

Department of Computer Networks
Polish-Japanese Institute of Information Technology
Warsaw, Poland

Abstract

All trust management systems must take into account the possibility of error: of misplaced trust. Therefore, regardless of whether it uses reputation or not, is centralized or distributed, a trust management system must be evaluated with consideration for the consequences of misplaced or abused trust. Thus, the issue of fairness has always been implicitly considered in the design and evaluation of trust management systems. This paper attempts to show that an implicit consideration, using the utilitarian paradigm of maximizing the sum of agents' utilities, is insufficient. Two case studies presented in the paper concern the design of a new reputation systems that uses implicit and emphasized negative feedbacks, and the evaluation of reputation systems' robustness to discrimination. The case studies demonstrate that considering fairness explicitly leads to different trust management system design and evaluation. Trust management systems can realize a goal of system fairness, identified with distributional fairness of agents' utilities. The realization of this goal can be achieved in a laboratory setting when all other factors that affect utilities can be excluded, and where the system can be tested using modeled adversaries. Taking the fairness of agent behavior explicitly into account when building trust or distrust can help to realize the goal of fairness of trust management systems.

Keywords: trust management, fairness, reputation, trust, Internet auction

1 Introduction

In distributed, open systems, where the behavior of autonomous agents is uncertain and can affect other agents' welfare, trust management is widely used to allow agents to determine what to expect about the behavior of other agents. Examples of practical use of trust management are (among others) reputation systems in online auctions and Peer-to-Peer file sharing systems.

This paper is concerned with the evaluation of trust management systems. Such systems must always take into account the possibility of error: of misplaced or abused trust. Therefore, the evaluation of trust management systems always should consider the consequences of such errors for the welfare or utility of users (autonomous agents that use the trust management system). The main claim of this

⋆ This research has been supported by the Polish Ministry of Science grant N N516 4307 33

[1] Email: adamw@pjwstk.edu.pl

doi:10.1016/j.entcs.2007.12.018

paper is that fairness should be explicitly taken into account in the evaluation of trust management systems. This claim is analyzed using two case studies that concern trust management systems that use reputation.

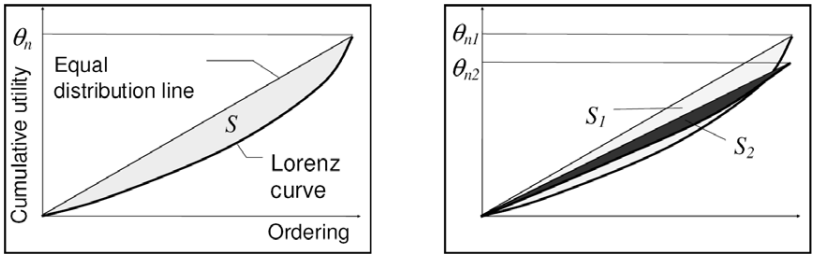
Reputation systems have usually been studied in one of two ways: by considering the real, economic or social impact of their practical applications (such as auction systems) [13] or by the use of simulation. New reputation systems are usually studied by simulation [3,8,9]. In this approach, the agent interaction is frequently modeled as a game, and a powerful paradigm prevails in such research. This is the utilitarian paradigm, originating from research on the Prisoner's Dilemma. Following the work of Axelrod [7], a large body of research has considered the emergence of cooperation (according to "Science", this is one of the 25 most important open scientific problems [2]). The introduction of reputation has been demonstrated as helpful to the emergence of cooperation .

In the Prisoner's Dilemma, the sum of payoffs of two agents is highest when both agents cooperate. This fact makes it possible to use the sum of payoffs as a measure of cooperation in the iterated Prisoner's Dilemma. This method is an utilitarian approach to the evaluation of reputation systems [3,14,16]. A reputation system is therefore considered successful (in most research) when the sum of utilities of all agents in the distributed system is highest. A notable exception is the work of Dellarocas [8]. Note that the utilitarian paradigm is used even if the simulation uses a more complex model of agent interaction than the Prisoner's Dilemma. A similar approach can be used in the evaluation of trust management systems that do not use reputation.

When considering the rationale of a trust management system using common sense, it seems that this approach is clearly inadequate. A trust management system that increases the sum of utilities of all agents does not consider whether, and to what extent, agents have been cheated (because of misplaced or abused trust). It is perfectly possible that a minority of agents is constantly being cheated, while the sum of utilities remains high. It is also possible that the sum of all utilities is high, but the variation of these utilities is also high. Evaluating trust management algorithms using the utilitarian approach disregards the fairness of all agents in the system. The reason for the adoption of this approach by previous research is not a definition of trust or trust management objectives that disregards fairness, but the lack of knowledge about computationally tractable methods for fairness evaluation.

This paper presents a case for the consideration of fairness of reputation and trust management systems. The main contribution of this paper lies in the presentation of a new, practical method for consideration fairness of trust management systems. The proposed method is based on a strong theoretical foundation: the theory of equity. The paper demonstrates that the proposed method can lead to different designs of reputation algorithms and that it is a useful criterion in the evaluation of reputation systems. In short: considering fairness is possible and can

[2] "Science", July 2005, special issue: "125 Questions: What Don't We Know?", http://www.sciencemag.org/sciext/125th/ However, note that the existence of reputation information is a modification of the original Prisoner's Dilemma. Axelrod has explicitly ruled out the existence of reputation information in his definition of the game.



The Lorenz Curve **Two incomparable distributions**

Fig. 1. Examples of Lorenz curves

make a difference for the design of trust management or reputation systems.

The fairness evaluation method described in section 2 can be used in simulation of any trust management system, given that the utilities of agents are defined. To demonstrate that the proposed method can be used in practice for consideration of fairness, two case studies are presented. The first study (section 3) concerns the evaluation of a new reputation algorithm, while the second study (section 4) is devoted to the robustness of a simple reputation algorithm to discrimination. The paper concludes with a discussion concerning the relationship of fairness and trust.

2 Evaluating Fairness Based on Theory of Equity

In order to consider fairness in the evaluation of trust management system, it is necessary to define fairness. In this section, the concept of fairness will be discussed and defined in an abstract manner. The next two sections will demonstrate how fairness can be used in practice to evaluate reputation systems and algorithms.

Fairness can be defined as a property of the behavior of individual agents or of a system. The behavior of agents can be called fair if it does not violate the contract, agreement or norms that govern the behavior of all agents in a system. The behavior of a system is fair if it promotes fair behavior of all agents in the system and results in a fair distribution of utilities of agents. In this work, the concept of system fairness is identified with distributive fairness, a sense narrower than social justice [10]. Distributive fairness is usually related to the question of distribution of some goods, resources or costs, be it kidneys for transplantation, parliament mandates, or the costs of water and electricity. Although extensively studied [11], fairness is a complex concept that depends much on cultural values, precedents, and the context of the problem. Therefore, a precise and computationally tractable definition is needed to use it in research.

Distributive fairness can be based on an axiomatic expression and can be expressed as a multicriteria optimization problem using theory of equity (this definition is also referred to as equitable optimization [12]). The axioms of the theory of equity which formulate a relation of preference on the outcome vectors of a multicriteria decision problem (these outcomes can be the utilities of agents in a distributed system). Let $y = [y_1, ..., y_n]$ be such an outcome vector (assuming there are n agents

that maximize their utilities). An equitable rational preference relation is any symmetric and transitive relation satisfying the following axioms (see [12] for formal definitions):

symmetry The ordering of the outcome values is ignored (e.g. a solution $y = [4, 2, 0]$ is equally good as a solution $y = [0, 2, 4]$).

monotony A outcome improving the value of one of the objectives is preferred, the values of other objectives are not deteriorated (e.g. $y = [4, 2, 0]$ is preferred to $y = [3, 2, 0]$).

principle of transfers A transfer of any small amount from an outcome to any other relatively worse-off outcome results in a more preferred outcome vector (e.g. $y = [3, 2, 1]$ is preferred to $y = [4, 2, 0]$).

The above axioms are sufficient to compare two outcome vectors on the basis of theory of equity, to determine whether they are equitably equivalent, or whether one of them is preferred to another. For instance, consider the two vectors: $y_1 = [5, 5, 10]$ and $y_2 = [2, 2, 2]$. It can be seen that by applying the axiom of monotony, vector y_1 is equitably preferred to y_2. This example demonstrates that the *theory of equity is not about equality*: vector y_2 has an equal distribution of outcomes, but it is considered worse, because in outcome y_1, all agents are better off. On the other hand, the two vectors: $[2, 4, 8]$ and $[2, 5, 6]$ are equitably incomparable: none of them can be preferred to the other.

However, the theory of equity can also be simply graphically described. The left part of Figure 1 shows one of the key concepts of distributive fairness: the Lorenz curve (to use its name from economics). The Lorenz curve is obtained by first taking the outcomes of all agents that participate in a distribution and ordering them from worse to best. Then, the ordered outcomes are added one at a time, creating a sequence of cumulative sums, called *cumulative ordered sums*. Let us denote this operation by a vector function $\theta(\overrightarrow{y}) = [\theta_1(\overrightarrow{y}), ..., \theta_n(\overrightarrow{y})]$ of the outcome vector. Then, the cumulative ordered sums of agents' utilities are calculated: starting from the utility of the worst agent (θ_1), then the sum of utilities of the worst and the second worst (θ_2), and so on, until the sum of all agents' utilities, which is denoted on the figure as θ_n. The second line on the figure, the equal distribution line, is simply a straight line connecting the points $(1, \theta_1)$ and (n, θ_n). The area between the two curves, denoted by S, can be seen as a measure of inequality of the agent's utilities. The objective of distributive fairness is to minimize this inequality, making the Lorenz curve as close to the equal distribution line as possible. Note that this objective frequently forms a tradeoff with the objective of maximizing the total sum of agents utilities (θ_n). The right part of Figure 1 shows two Lorenz curves that correspond to different distributions among the same agents. The first distribution has a higher θ_n, but also a higher inequality, while the second distribution has a lower total of agents' utilities, but is more fair. In terms of equitable optimization, the two distributions on the right side of Figure 1 are incomparable - the choice of one of them depends on the preferences of a decision maker.

Without a detailed discussion, let us state that the theory of equity allows for the formulation of fairness objectives as a multicriteria optimization problem, and

that this transformation relies on the operation θ of ordering and then creating cumulative ordered sums. (The Pareto-optimal solutions of the multicriteria problem obtained by performing cumulative ordered sums of the criteria of the original problem are equitably optimal solutions of the original problem.) Therefore, the Lorenz curve actually represents the criteria that need to be optimized in order to achieve equitably efficient (fair) solutions. These criteria are the cumulative sums of $1, 2, ..., k$ worst outcomes. Note also that according to the theory of equity, the utilitarian approach is a special case: optimizing the sum of agent's utilities also leads to an equitably efficient solution. However, this solution is one of many and it is up to the decision maker (the designer of a trust management system) to make a choice of one solution. At the other extreme, the solution obtained by equalizing the outcomes of all agents is also equitably efficient. The interested reader is referred to [12].

The area between the Lorenz curve and the equal distribution line can be simply calculated and used as a computable measure of inequality. It can be shown that minimizing this measure leads to fair distributions [12]. The Gini coefficient (frequently used in economics) is the area S normalized by θ_n: $Gini = \frac{2S}{\theta_n}$. Note that minimizing the Gini coefficient to obtain fair distributions can lead to worse total outcomes (sums of all agent's utilities) - this drawback can be overcome when the problem of distributive fairness is formulated as a multicriteria problem. Solving such a problem means finding a set of Pareto-optimal solutions and choosing one of them that best suits the preferences of a decision maker (which are expressed as additional input to the problem). The decision maker can then choose whether he prefers increasing equality at the cost of decreasing the total outcome, or not.

Such a definition can be applied in the study of reputation or trust management systems in a laboratory setting, where other factors that affect agent's utilities can be excluded because they influence all agents equally (this is equivalent to the ceteris paribus assumption used in economics). In a real environment, a trust management system would interact with agents that have diverse resources, abilities and motivation, and therefore the definition of fairness would become much more difficult. However, the study of trust management systems usually begins in a laboratory where the conditions required for consideration of fairness can be satisfied.

3 First Case: a New Reputation Algorithm

The first case study that demonstrates the consideration of fairness in the design and evaluation of trust management systems is devoted to a new reputation algorithm that has been developed for online auctions. The algorithm is an extension of the simple algorithm used by eBay by taking into account a new type of feedback: the so called "*implicit feedback*" [9], and additionally the algorithm is capable of stressing the importance of explicit negative feedbacks. The consideration of fairness has influenced the design of this algorithm.

3.1 The algorithm of implicit and stressed negative feedbacks

Most online auction sites use a simple feedback-based reputation system [13]. Typically, parties involved in a transaction mutually post feedbacks after the transaction is committed. Each transaction can be judged as '*positive*', '*neutral*', or '*negative*'. The reputation of a user is simply the number of distinct partners providing positive feedbacks minus the number of distinct partners providing negative feedbacks (possibly normalized by the number of all distinct partners). As pointed out in [14], such a simple reputation system suffers from numerous deficiencies, including the subjective nature of feedbacks and the lack of transactional and social contexts. Yet another drawback of feedback-based reputation systems is that these systems do not account for psychological motivation of users. Many users refrain from posting a neutral or negative feedback in fear of retaliation, thus biasing the system into assigning overestimated reputation scores. This phenomenon is manifested by high asymmetry in feedbacks collected after auctions and, equally importantly, by high number of auctions with no feedback provided. Many of these missing feedbacks may convey implicit and unvoiced assessments of poor seller's performance which should be included in the computation of a seller's reputation.

As described in [9], there can be many ways of identifying implicit feedbacks in a real-world reputation system, based on the observation of behavioral patterns. To evaluate the effectiveness of using implicit feedback, we have identified a simpler reputation algorithm that can be simulated and compared to the algorithm of most Internet auction houses.

If a user decides to post a negative feedback in spite of the negative incentives involved, it is probable that the user is very dissatisfied with his contractor. For that reason, his negative feedback should perhaps have a higher impact on reputation than an ordinary positive feedback. We shall refer this method as stressing of negative feedbacks.

Consider a user u with a history of n auctions. Let us assume that only $m \leq n$ of these auctions have a feedback. Out of these m feedbacks m^+ are positive or neutral feedbacks (in practice, the amount of neutral feedbacks can be ignored), m^- are negative feedbacks, while $m^* = n - m$ is the amount of missing feedbacks (transactions that had no feedback). Thus, $m^+ \leq m \leq n$. The reputation ρ_u of the user u will be calculated as follows:

$$\rho_u = \frac{m^+}{\alpha m^* + m^+ + (1+\beta)m^-}$$

where $0 \leq \alpha, \beta \leq 1$. Thus, if $\alpha = \beta = 0$, the above reputation score becomes a simple ratio of the number of positive feedbacks received by the user u. In the case when the user has had no auctions, the above formula is undefined. In such case we set the reputation ρ_u to an initial value, ρ_0. The two coefficients have the following complementary roles: α is used to control the importance of implicit negative feedbacks, and β controls the stress that is placed on explicit negative feedbacks.

To be precise, in our simulations we use a slightly more complex version of the above algorithm. Since agents in the simulator choose whom they want to

interact with on the basis of reputation scores, it is necessary to avoid that the reputation would drop suddenly to a low level. This can happen in the initial phase of the simulation, when the reputation score has not yet stabilized (initially, a single negative feedback could decrease the initial reputation by a large degree). Therefore, we use a simple moving average to smooth reputation changes. The smoothed reputation $\rho_u^{ma}(t) = 0.5\rho_u^{ma}(t-1) + \rho_u(t)$, where t is time, and $\rho_u^{ma}(0) = \rho_0$(the smoothed reputation is initialized by the initial reputation value). Note that over time, the impact of the initial reputation decreases exponentially.

3.2 The simulator

In the design of the simulator, we had to make a decision about a sufficiently realistic, yet not too complex model of the auction system, of user behavior, and of the reputation system. We chose to simulate the reputation system almost totally faithfully (the only simplification is that we use only positive and negative feedbacks). The behavior of a user is also quite realistic: the user can take into account reputation when choosing a business partner. The user also decides whether she wishes to report or not, depending on the type of feedback. Users can cheat in feedbacks, as well as in transactions. Users may also use transaction strategies that depend on the history of their individual interactions, as well as on the reputation value of the other participant.

The auction system, on the other hand, has been simplified. We consider that it was not necessary to simulate the entire auction process; rather, we simulate the selection of users using random choice of a set o potential sellers. The choosing user (the buyer) selects one of the sellers that has the highest reputation in the set. The transaction itself has also been simplified: the simulator could use the Prisoner's Dilemma or a zero-sum game.

In our simulator, a number of agents that represent users interact with each other. The reputation system is maintained by a reputation server that is also used to summarize the outcomes of agent interactions (user transactions). Each agent is described by the following parameters: r^+, the probability that an agent will send a feedback if it is positive; r^-, the probability that an agent will send a feedback if it is not positive; the chosen game strategy; the reputation threshold ρ_{min} that is used by some strategies; and the probability of cheating c that is used by some strategies. We can specify a number of agents and every agent can have distinct parameters. However, we usually partition all agents into two sets that have the same parameters, called the honest and dishonest agents.

The two game strategies used in the simulations described in this paper are: to cheat with the probability c, or to play Tit-for-Tat with a reputation threshold ρ_{min}. Tit-for-tat is a famous strategy for the iterated PD game that has been proposed by Rapaport. This strategy works simply by repeating the move made by the other agent in a previous encounter. If two agents meet for the first time, the classic Tit-for-tat strategy always cooperates, allowing the agents to start an unending pattern of honest transactions. We modify Tit-for-tat to use a reputation threshold: if two agents meet for the first time and the second agents' reputation is below ρ_{min}, the

first agent defects.

The server can compute reputations using all available feedbacks and using any implemented algorithm. The results of the simulation include: the reputations of individual agents and the total payoffs (from all transactions) of all agents. The payoffs are affected by the way the reputation system works: for example, if agents send very few feedbacks, reputations will be random and the payoffs of good agents will drop. The simulator allows to check whether the implemented reputation algorithms are effective.

To verify the concept of implicit and stressed negative feedbacks, we have implemented the reputation algorithm described above. User feedback behavior was simulated in one of two ways: either all agents always posted feedback truthfully (*perfect feedback*), or agents posted feedback according to the following rule (*poor feedback*). If the feedback was positive, an agent would post it with probability $r^+ = 0.6$, and if the feedback was not positive, the agent would post it with probability $r^+ = 0.05$. All feedbacks were always true, if they were posted. The parameters of poor feedback behavior were observed from the analysis of traces obtained from a Polish Internet auction house. The traces contained over 650000 committed auctions made by a sample group of 10000 buyers over a period of six months. While it is clear that the posting behavior of real users is more complex, this approximation is sufficient to evaluate the possibility of using implicit feedback in a simple reputation algorithm.

The algorithm was evaluated using the following simulation scenarios. There were 300 simulated agents that were divided into two sets: the *honest agents* and the *dishonest agents*. Honest agents were 90% of all agents, and the remaining agents were dishonest. An honest agent used the Tit-for-Tat strategy with a reputation threshold of $\rho_{min} = 0.5$. A dishonest agent used a strategy of random cheating with probability $c = 0.6$. All agents used poor feedback or perfect feedback, depending on the scenario. In all simulations, 40000 auctions were simulated.

Together, there have been three significant simulation scenarios: perfect feedback with reputation calculated using a simple ratio of positive feedbacks (a reputation algorithm like described in the previous section, only with $\alpha = \beta = 0$); poor reports with a simple ratio; and poor reports with the reputation algorithm that uses implicit and stressed negative feedbacks, with different settings for α and β.

All experiments were conducted using the Monte-Carlo method. We present average results from 10 simulation runs, together with 95% confidence intervals of results (intervals for which with probability of 95%, all results would belong to the interval. Confidence intervals were obtained using the t-Student distribution). The outcomes of every simulation were the average payoffs and the Gini coefficients of honest and dishonest agents.

3.3 Considering fairness in algorithm evaluation

We evaluate the effectiveness of a reputation system using the following criteria: the average payoff of a good agent, the average payoff of a bad agent, and the Gini coefficient of the payoffs of the good agents. The last criterion was introduced as a

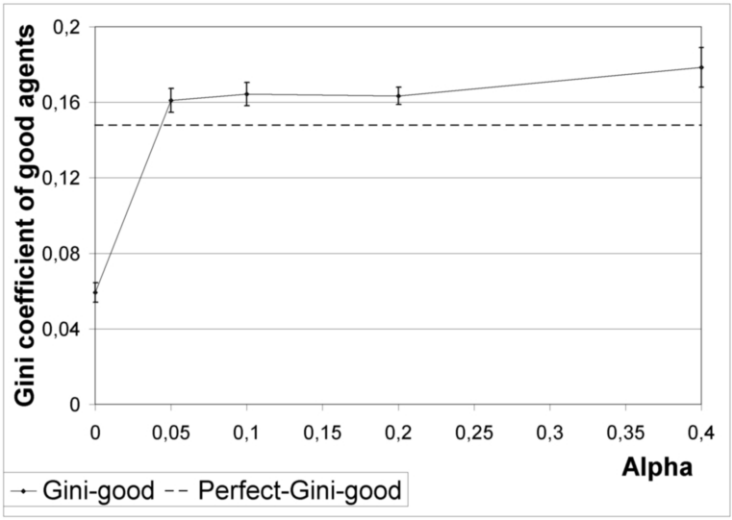


Fig. 2. Average Gini coefficient of honest agents

way of evaluating the effectiveness of the reputation system in providing fairness of the treatment of good agents.

The objective of the evaluation was to choose a good value of the parameters α and β for the reputation algorithm. Choosing low values of α and β would result in an algorithm that behaved almost like the simple ratio of positive feedbacks. Choosing high values of α would decrease the reputation of agents that persistently had many missing feedbacks. The effect of these implicit negative feedbacks could become stronger than the ratio of positive feedbacks, and could decrease the reputations of users too much (bias the reputation in another direction). On the other hand, increasing β may become dangerous if agents can send false negative reports.

The results of algorithm evaluation are shown on Figures 2 and 3. The figures show results in the poor feedback scenario for different values of α: $0, 0.05, 0.1, 0.2$, and 0.4, and for $\beta = 0.05$. The effect of increasing β will be described later - first, let us consider the effect of varying α .

Figure 2 shows the effect of varying α on the Gini coefficient. Each point shows the average Gini coefficient and its confidence intervals. Figure 3 shows the payoffs of good (top curve) and bad agents (bottom curve). Both figures also show results for perfect feedback that were used as reference levels. These values are shown as straight dashed horizontal lines labeled "perfect".

Values obtained for $\alpha = 0$ show the performance of a simple reputation algorithm that does not use implicit feedback (but slightly stresses explicit negative feedbacks) under the poor feedback scenario. Recall that the parameters for the poor feedback scenario correspond to real-world data, and therefore these results approximate a realistic performance of an algorithm such as used by most Internet auction houses. Comparing these results to the perfect reference levels, it can be noticed that while the average payoff of honest agents decreases slightly, the average payoff of dishonest agents increases by over 100%. *This is an indication of how easy it is for dishonest agents to exploit the poor quality of feedbacks in a simple reputation system.* These

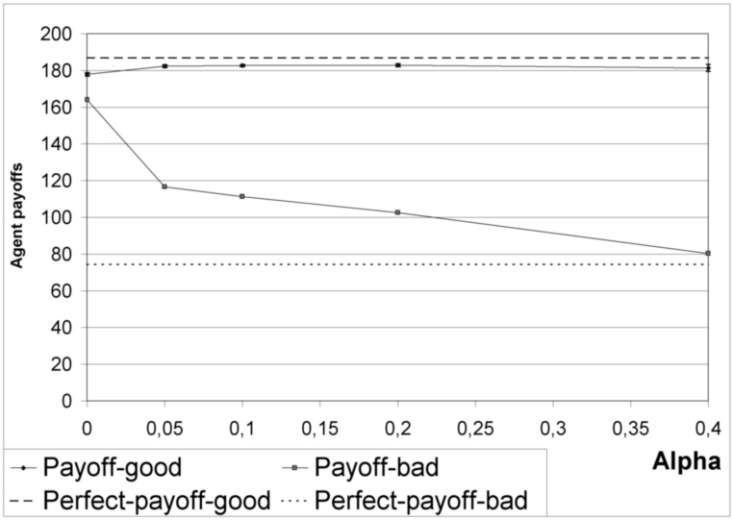


Fig. 3. Average payoffs of honest and dishonest agents

results also show that the average or total payoff of good agents can be a poor indicator of the volume of fraudulent transactions.

Increasing the value of α demonstrates the effectiveness of using implicit feedback. Even for very small $\alpha = 0.05$, the average payoff of dishonest agents drops dramatically, almost back to the "perfect" reference level. The average payoff of honest agents increases slowly with the increase of α. The average payoff of dishonest agents decreased continually with the increase of α, dropping almost to the level obtained for perfect feedback. The average payoff of honest agents increased for $\alpha < 0.3$; for higher α, the average payoff of honest agents began to decrease slightly. Table 1 shows 95% confidence intervals for the average payoffs of honest agents (row *PHCI*) and for values of the Gini coefficient of all honest agents (row

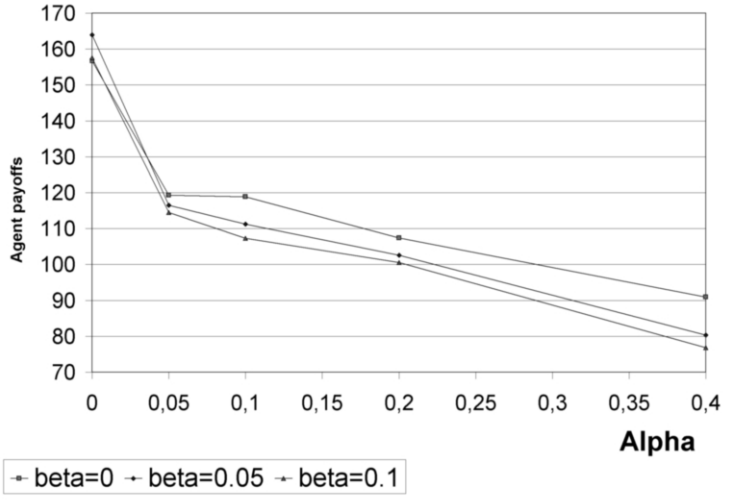


Fig. 4. Effect of stressing negative feedbacks'

	P	0	0.05	0.1	0.2	0.4
PHCI	186,2-187,4	176,9-178,8	181,6-183,1	181,9-183,3	182,1-183,6	179,4-183,2
GHCI	0,14-0,15	0,05-0,06	0,15-0,17	0,16-0,17	0,16-0,17	0,17-0,19

Table 1
95% confidence intervals for payoffs and Gini coefficients

GHCI). The confidence intervals were obtained from samples of 10 simulation runs. The same confidence intervals are also shown on the figures in the form of vertical lines around the data points. The column labeled P describes the perfect feedback scenario, while the other columns describe the poor feedback scenario with different values of α and for $\beta = 0.05$. Note that the average payoff of honest agents changes very little with the variation of α above 0.05. The Gini coefficient is more sensitive to changes of α. The sharp increase of the Gini coefficient when α is increased from 0 to 0.05 is due to the fact that when implicit feedbacks are not taken into account, honest agents are worse off, but the distribution of agents utilities can be more even. The Gini coefficient should therefore not be used as a single criterion, but rather jointly with the total utility or a similar criterion.

On the basis of the average payoffs alone, a designer of the reputation system would choose a value of $\alpha = 0.2$, or even $\alpha = 0.4$. Larger values of α lead to an increased penalty of dishonest agents, and the average payoff of good agents is largest for $\alpha = 0.2$.

However, consideration of the Gini coefficient shows clearly that increasing α leads to an increased inequality in the treatment of agents. This may be due to the fact that agent reputation becomes more variable due to the additional effect of missing feedbacks that depends on random variables. Reputation is no longer dependent on agent behavior alone. The optimal value of α, when considering the Gini coefficients alone, is $\alpha = 0$. Unfortunately, this value results in very poor levels for the other criteria. Therefore, a system designer that would consider all criteria jointly would probably choose $\alpha = 0.05$.

The effect of increasing β is demonstrated on Figure 4. The figure shows again the relationship of average payoffs of bad agents, and α. However, the figure shows three lines that have been obtained from varying β. As β increases, the average payoff of bad agents drops for all values of α. The effect is apparent for small values of β; as β exceeds 0.1, it stops having the effect of decreasing payoffs of bad agents, while at the same time it strongly increases the Gini coefficient of good agents. Also, the risk of exploiting false explicit negative feedbacks for coalition attacks increases with higher values of β. Therefore, we have limited our study to small values of β.

Concluding, the first case study demonstrates the effectiveness of a new reputation algorithm that uses implicit and stressed negative feedbacks. However, implicit and stressed negative feedbacks are a dangerous tool that must be used carefully. Increasing their weight in a reputation algorithm may result in penalizing honest agents along with dishonest agents, or in increasing the risk of coalition attacks. This deficiency is most clearly demonstrated when one considers the distributive fairness of honest agents' payoffs in a laboratory setting, where all other factors that influence payoffs can be excluded.

	Discrimination	Cooperation
Average payoff of old agents	86612	84753
96% confidence interval	82223-91001	83864-85642

Table 2
Old agents' payoffs in the first simulation

4 Second Case: Robustness of Reputation Systems

The second case study concerns the evaluation of a simple reputation system's resistance to discrimination attacks. In research on reputation systems, the adversary model usually permits collaboration of adversaries. This collaboration frequently takes the form of a coalition that attempts to influence reputation scores. Discrimination is another form of adversary behavior; the adversary need not form an explicit coalition with a single goal, but is allowed to distinguish agents belonging to two or more groups. The adversary can treat one agents from one group fairly, while cheating other agents. Discrimination has been studied, among others, by Dellarocas [8].

In this case study we consider an open market, where a number of agents trades some goods. The majority of agents on the market belong to one group, called the *old agents*. Newcomers constantly enter the market, but they usually form a minority group, called the *new agents*. Old agents usually trade fairly with other old agents, but they also usually cheat new agents. New agents usually act fairly (at least initially). They can leave the market any time if their losses are too high, or - after a time - they may join the group of old agents.

We have chosen to simulate a steady state of the described market, where a constant proportion of agents are new agents. The processes of entering and leaving the market and of joining the group of old agents have been left out of the simulation. It has been our goal to answer the following questions: is discrimination profitable for old agents? Can the reputation system protect against discrimination? And how can we evaluate the effectiveness of the reputation system in preventing discrimination?

In our case study we have used the same simulator as described in the previous section. We have simulated 100 agents, out of which 30 were new agents. The agents executed 20000 transactions that have been modeled using the Prisoner's Dilemma. Therefore, the maximum total payoff of all agents was equal to 120000. As in the previous study, the presented results were obtained from 10 simulation runs. Unlike

	New agents	Old agents
Average reputation	30	75
95% confidence interval	26-34	72-78

Table 3
Reputations of old and new agents under discrimination

	Discrimination	Cooperation
Average total payoff	107848	107833
Average Gini	0,65	0,53

Table 4
Total payoffs and the Gini coefficient in the second simulation

the previous study, we have always simulated perfect feedback behavior.

In the first simulation, there were two scenarios. In the *cooperation scenario* all agents (new and old) always cooperated with each other. The results of this scenario are used as reference levels for comparison. In the *discrimination scenario*, all new agents used the Reputation Tit-for-tat strategy, and all old agents used a new strategy of discrimination: they always cooperated with old agents and always cheated new agents. The results of the first simulation enabled us to answer the first two questions. The average total payoffs of old agents are presented on Table 2. The distribution of total payoffs was shifter towards the larger end of the confidence interval.

In this case, old agents are indeed better off using discrimination: in other words, it sometimes pays off to cheat.

What happened to the reputation system? Table 3 shows the average reputation of new and old agents under the discrimination scenario. Paradoxically, the reputation of new ("honest") agents is only 30, while the reputation of old ("cheating") agents is as high as 75. This is explained by the fact that when new agents retaliate against cheating old agents, they receive negative feedback. All agents used perfect feedback, and the new agents were a minority forced to retaliate against a majority of old agents. On the other hand, old agents only cheat a minority and maintain a high reputation by cooperating with each other. In this case, the reputation system did not protect honest agents; it acted against them.

The first simulation used an extreme case when the detection of discrimination is simple. Old agents always discriminate against new agents, and as a result, the total payoff of all agents decreases noticeably. However, discrimination can be more subtle. What if old agents discriminate only sometimes? What if new agents cheat sometimes, too? In such a case, how do we evaluate whether discrimination was prevented by the reputation system?

To answer the question of how to detect discrimination and how to evaluate the effectiveness of the reputation system in preventing discrimination, the second simulation used two scenarios. In the discrimination scenario, old agents always discriminate; but in the second scenario of *no discrimination*, all agents cheat with the same probability of $c = 0.24$. The results are summarized on Table 4. It turns out that using the total payoff of all agents, the two scenarios are almost indistinguishable (the difference in total payoffs is about 0.01%). On the other hand, the Gini coefficients of all agents' payoffs differ widely in the two scenarios. This is due to the fact that under the discrimination scenario, the distribution of agents' payoffs is much less fair than under the scenario of no discrimination.

The second case study demonstrates that when adversaries are allowed to discriminate other agents, fairness becomes an even larger issue in reputation systems. Taking fairness into consideration may also help to evaluate the extent of discrimination in a laboratory setting, and to determine whether a reputation system effectively protects against discrimination.

5 Conclusion

The two case studies have demonstrated that *fairness can be taken into account in the evaluation of trust management systems in a laboratory setting. Considering fairness leads to different design of reputation algorithms, and fairness is a useful criterion in the evaluation of reputation systems.* However, the consideration of fairness is based on a premise that whenever possible, trust management systems should be designed to provide fair treatment of their users. We would like to go one step further and open a discussion of the questions: should fairness be a goal of trust management systems? If so, how can this goal be achieved in practice?

The first question raises some concerns. In a real-world setting, users of trust management systems would be expected to have quite varied levels of utility (perhaps even incomparable ones). How, then, do we expect a trust management system to realize a goal of fairness?

This concern is based on a frequent misconception that mistakes equality for fairness. If a trader in an Internet auction house has better goods, provides better services and has better marketing than other traders, it is perfectly fair that he should have a larger transaction volume and a larger revenue. In fact, his reputation should increase, as well, so the trust management system should in this case support him in getting even more trade. On the other hand, if we have two traders that have comparable goods, services, and marketing, yet they have very unequal reputation and transaction volumes, surely something is wrong in the way the trust management system works.

Therefore, *when all other factors can be excluded, fairness can be identified with distributional fairness and the concept of equity.* In a laboratory setting, such conditions can be satisfied and we can design trust management systems that realize the goal of fairness, even in the presence of adversaries.

The *theory of equity allows for the existence of solutions that satisfy multiple constraints*, and *are considered equitably efficient if no other solution dominates them* in terms of the equitable preference relation. Therefore, even when agents' utilities are the result of their operations on a market with different resources and capabilities, but all agents *act fairly*, then the resulting distribution of utilities may be equitably efficient.

Yet the question remains how fairness can be achieved by a trust management system in a realistic environment where agents are not expected to act fairly. The two concepts of fairness, individual agent fairness and fairness of a trust management system, seem to be related. If the trust management system will increase trust as a direct consequence of fair agent behavior, it should be able to realize its own

system goal of fairness. If traders in an Internet auction house can only increase their reputation as a result of fair behavior (for example, if all transactions are supervised by impartial, omniscient arbiters), then the trust management system will be perfectly fair - the distribution of agent utilities should be affected only by legitimate factors. Of course, the real challenge lies in the design of such trust management systems.

Therefore, *the fairness of a trust management system in a realistic setting should be an emergent property that depends on the fairness of individual agents.* The trust management system should provide incentives for fair agent behavior. Notice that this is also possible in the case of a trust management system that does not use centralized control and global information. For example, a Peer-to-peer file sharing system that combats free-riding using trust management can also aim for fair treatment of users, while it is impossible in such a system to maintain global information and use centralized control. Global information is only used in a laboratory for the evaluation of such a system. The system itself works based on local information only.

Yet the main goal of trust management systems is the support of decisions under uncertainty on the basis of justified trust. While this goal, and the previous definitions of trust, *are not contradictory to the consideration of fairness, they have not before considered fairness explicitly.* It seems necessary to go on to the next question: what is the relationship between trust and fairness?

Let us recall the two main (simplified) definitions of trust used in literature:

Definition 5.1 Trust is a subjective, context-dependent expectation that an agent will carry out a particular action in a situation of uncertainty. [2,3]

Definition 5.2 Trust is the degree of dependence on another agent that the trustee is willing to accept in a situation of uncertainty, considering risks and incentives involved. [4,5]

It has been pointed out in [1] that trust has yet another dimension: that of norms, values, or principles of human agents. If an agent perceives that another agent has high principles and acts in accordance with social norms, she is more likely to trust him. This observation shows that trust can be affected by the fair behavior of agents.

Elgesem's observation can be extended to non-human agents, because in many cases, individual fairness can have computational models. This is the case when the normative behavior of agents can be specified using contracts [15,16]. While in some application domains expressing contracts might be difficult, in other domains it is quite simple. Consider the case of electronic games: there, fair behavior is simply behavior that does not violate game rules. Trust management systems can be used in such an application [17] and it is quite clear that the level of trust directly depends on the fairness of an agent.

The exact relationship of trust and fairness requires further study in the areas of psychology and social science. However, it seems that the question deserves more attention from researchers in the field. An interesting questions is whether ensuring

fairness eliminates the need for trust. In other words, can a trust management system be renamed as a "fairness management system"? Is it possible to achieve the goals of trust management without the use of trust? If not, then how should trust be defined if the goal of the trust management system is providing fairness? Without claiming to have resolved the issue, we would like to open a discussion by proposing two revised definitions of trust:

Definition 5.1a. Trust is a subjective, context-dependent expectation that an agent will behave fairly with respect to my interests in a situation of uncertainty.

Definition 5.1b. Trust is the degree of dependence on another agent that the trustee is willing to accept in a situation of uncertainty, considering risks, incentives and the fairness of the trusted agent.

The issue of fairness has always been implicitly considered in the design and evaluation of trust management systems. This paper has attempted to show that an implicit consideration, using the utilitarian paradigm, is insufficient. Considering fairness explicitly leads to different trust management system design and evaluation. Trust management systems can realize a goal of system fairness, identified with distributional fairness of agents' utilities. The realization of this goal can be achieved in a laboratory setting when all other factors that affect utilities can be excluded, and where we can test the system using modeled adversaries. In a realistic setting, fairness may be an emergent property of trust management systems. In order to support its emergence, the relationship between trust and fairness requires closer study.

References

[1] Elgesem, E., *Normative Structures in Trust Management*, Trust Management (iTrust 2006), Springer, LNCS 3986, 2006

[2] Gambetta, D., *Can We Trust Trust?*, in D. Gambetta, editor, "Trust: Making and Braking of Cooperative Relations", Basil Blackwell, Oxford, 1988

[3] Mui, L., Computational Models of Trust and Reputation: Agents, Evolutionary Games, and Social Networks, Ph.D. Dissertation, Massachusetts Institute of Technology, 2003

[4] Josang, B. C. Keser and T. Dimitrakos, *Can We Manage Trust?*, Trust Management (iTrust 2005), Springer, LNCS 3477, 2005

[5] Marsh, S., P. Briggs and W. Wagealla, *Enhancing collaborative environments on the basis of trust*, 2004

[6] Luhmann, N., *Familiarity, confidence and trust: problems and alternatives*, in D. Gambetta, editor, "Trust: Making and Breaking of Cooperative Relations", Basil Blackwell, Oxford, 1988

[7] Axelrod, R., The Evolution of Cooperation, Basic Books, New York. 1984

[8] Dellarocas, C., *Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior*, In Proc. of the 2nd ACM Conference on Electronic Commerce, Minneapolis, MN, USA, October 17-20, 2000

[9] Wierzbicki, A. and M. Morzy, *The Sound of Silence: Mining Implicit Feedbacks to Compute Reputation*, Proc. 2nd international Workshop on Internet and Network Economics (WINE'06), Springer, LNCS, 2006

[10] Rawls, J., The Theory of Justice. Harvard Univ. Press, 1971.

[11] Young, H. P., Equity: In Theory and Practice. Princeton University Press, 1994.

[12] Ogryczak, W., M. M. Kostreva and A. Wierzbicki, *Equitable aggregations and multiple criteria analysis*, European Journal of Operational Research, 158:362-377, 2004

[13] Resnick, P. and R. Zeckhauser, *Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system*, Advances in Applied Microeconomics, 11, 2002.

[14] Malaga, R. A., *Web-based reputation management systems: Problems and suggested solutions*, Electronic Commerce Research, 4 vol. **1**, 2001.

[15] Tan, Y., W. Thoen and J. Gordijn, *Modeling controls for dynamic value exchanges in virtual organizations*, Trust Management: Second International Conference, iTrust 2004, Oxford, Springer, LNCS 2995, pp. 236-250, 2004

[16] Gordijn, J. and H. Akkermans, *Designing and evaluating e-Business models*, IEEE Intelligent Systems, vol. **16**, pp. 11-17, 2001

[17] Wierzbicki, A., *Trust Enforcement in Peer-to-peer Massive Multi-player Online Games*, Proc. Grid computing, high-performAnce and Distributed Applications (GADA'06), Springer, LNCS, 2006