ICORE'09

international conference on reputation.theory and technology

Proceedings of the First International Conference on Reputation

18-20 March 2009
Gargonza, Italy

How to cite this proceedings book?

```
@proceedings{ICORE09,
         Title = {Proceedings of the First International Conference on Reputation:
                  Theory and Technology - ICORE 09, Gargonza, Italy},
         Editor = {Mario Paolucci},
         Isbn = {978-88-85059-27-6},
         Url = {http://www.reputation09.net},
         Month ={March},
         Year = {2009}}
```

Editor: Mario Paolucci

CNR - National Research Council
Roma, Italy

ISBN 978-88-85059-27-6
9 788885 059276

Introduction

This volume presents the pre-proceedings of ICORE 2009: The First International Conference on Reputation: Theory and Technology. The ICORE conference intends to provide a ground for the scientific discussion on reputation systems, with a perspective on policy and e-Government. Reputation is an ancient artifact; its origins are connected with the origins of language itself - even pre-dating it in the famous interpretation of gossip by Robin Dunbar. The role of reputation has evolved with human society, from an efficient solution to the problem of grooming in large groups, to a mechanism supporting norm compliance, a tool for in-group signaling, a weapon of large collectives to exhalt or calumny public figure; the older roles blending with the newer ones, giving raise to a multi faceted artefact.

In the social and economical sciences, the role of reputation as a partner selection mechanism started to be appreciated for cooperation in the early eighties. Despite important advances in the study of cooperation networks, still no explicit theory of the cognitive ingredients and processes which reputation is made of was provided. In current usage, reputation is still - with few exceptions - viewed simply as an attribute in the decision for partner selection.

More recently, reputation and gossip started to become crucial in many fields, for example organisation science and management, governance, business ethics, etc. where the importance of branding became visible. In these domains, reputation has soon become an intangible asset. The economic reading of the issue at hand implied an extension of reputation to super-individual levels, requiring a still wanting conceptual clarification and interdisciplinary investigation.

In addition, reputation is increasingly at the centre of attention in many fields of science and domains of application, including but not reduced to policy-making, (e-)governance, cultural evolution, social dilemmas and socio-dynamics. Even so, there is a great deal of ad hoc models, and little integration of instruments for the implementation, management and optimisation of reputation. On one hand, policy makers, entrepreneurs and administrators deem it possible to manage public, corporate and firm reputation without contributing to or accessing a solid, general and integrated body of scientific knowledge on the subject matter. On the other hand, software designers believe they can design and implement online reputation reporting systems without investigating what the properties, requirements and dynamics of reputation in natural societies and why it did evolve for.

Reputation, instead, deserves a full role as a scientific topic, with focus on its specificities, as ist potential as preventive social knowledge, and the selective mechanism of transmission. Only by recogizing the central role of reputation and its cross disciplinary nature we could obtain the advances that are not even pursued inside the borders of a single discipline.

In the view of the organisers of this conference, reputation is an old artefact for answering a new challenge, and that is the regulation of complex, global, electronic societies. Innovation demands that the potential of old instruments are fully understood and exploited, in order to be incorporated into novel, intelligent technologies. We invited contributions from all the fields of reputation science, from philosophy to experimental economics, from game theory to computer science and to sociology. Of thirty two submissions, we present in these proceedings fifteen papers.

The papers collected in these proceedings have been grouped in four sections: Foundations of Reputation, Field and Laboratory Research on Reputation, and Simulation of Reputation.

In the Foundations section we have contributions from computers science, artificial intelligence, management science and sociology; two papers (Sierra and Debenham, Garcin et al.) touch the critical point of aggregation of reputation feedbacks, the first from a more abstract point of view, the second making explicit reference to current web 2.0 systems. The paper from Niemann et al. developes a game theoretical model of the eBay scoring system, providing

some explicit claims on the effects of recent policy changes. Social networks, the medium that allows for reputation circulation, which in turn may cause a reshaping of the network itself, are discussed in the paper from Corten and Cook. Vague, uncertain, or untestable assertions are often associated with gossip and reputation. The paper from Kramer and Rosenthal proposes a contribution on this fondamental issue.

The section on Field and Laboratory Research on Reputation contains several contributions. In a paper on perceived informal power in organizations from Labun et al. explore the link between power and trust. In a laboratory experiment on trust, Buskens et al. compare a condition in which the trustor knows the result of other games played by the trustee with one in which she does not.

A group of reseachers from sociology and economy departments contributes with two papers, both of which have as first author Riccardo Boero. In the first paper, they interpret the behavioral patterns of subjects taking a decision on economic investment and reprodces these patterns in an agent-based model. In the second, they introduce a third-party rater in the trust game, taking reputation more explicitly into account. With the work of Bapna et al, we move back into Internet territory, exploring online auctions with a focus in simultaneous auctions for identical goods. All these field experiments provide the essential micro data to validate models and theories of reputation.

Coming to Simulation of Reputation, we start the section with two contributions on peer to peer networks, both focused on the issue of reputation systems as a mean to fight malicious users who spread inauthentic files to gain credits. In the first paper, Boella and collaborators show how reputation systems are effective only if there is a widespread cooperation by users in verifying authenticity of files starting during the download phase; Rosas and Bonnaire introduce the notion of risk associated to the reputation value, arguing that reputation and risk together work better than reputation alone.

The paper from Krupa et al. propose two scenarios to improve the relevance of the ART testbed, a computer science competition aimed at selecting the best trust strategies, currently used as a benchamrk for trust algorithms. With the contribution of König et al., we see reputation at work in a new, challenging field, that of the Internet of Services, where several kind of providers and service buyers compete in a market characterized by standardized and low-value transactions. Finally, Quattrociocchi and myself close the proceedings with an application of the Repage reputation for reputation and image management, applied to a simple market with growing levels of informational cheating.

Acknowledgements

This book has been made possible with financial support from the European Community under the FP6 programme (eRep project, contract number CIT5-028575).

The organizers of this conference would like to thank the scientific committee members and reviewers, the authors of the submitted papers, the Gargonza castle staff for the great atmosphere, and all those who have helped.

Organization

The First International Conference on Reputation: Theory and Technology (ICORE 2009) was sponsored by the eRep under the 6th Framework (Contract n. 28575) and organized by *Institute of Sciences and Technologies of Cognition* - CNR, Italy; *University of Groningen, Marketing Institute* - Netherlands; *CSIC-IIIA Institut d'Intestigació en Intel·ligència Artificial, Spanish Scientific Research Council*, - Spain; and *UBT University of Bayreuth, Dept. of Information Systems* - Germany.

Scientific Chair

Mario Paolucci

Local Organizers

Federica Mattei
Stefano Picascia
Walter Quattrociocchi

Program Commitee

Frederic Amblard
Alvaro Arenas
Lucio Biggiero
Vincent Buskens
Cristiano Castelfranchi
Claudio Cioffi-Revilla
Helder Coelho
Chris Dellarocas
Gennaro Di Tosto
Bruce Edmonds
Torsten Eymann
Boi Faltings
Liane Gabora
Francesca Giardini
Nigel Gilbert
David Hales
Rainer Hegselmann
Tomas Klos
Wander Jager
Sverker Janson
Marco Janssen
Audun Josang
Ioan Alfred Letia
Churn-Jung Liau
Paolo Massa
Daniel Moldt
Mario Paolucci
Stefano Picascia
Daniele Quercia
Werner Raub
Jordi Sabater-Mir
Jean-Marc Seigneur
Sandip Sen
Jaime Sichman
Carles Sierra
Luís Silva
Munindar Singh
Flaminio Squazzoni
Klaus Troitzsch
Kerstin Voß
Ramin Yahyapour
Christof Weinhardt

Referees *(not included in the Program Commitee)*

Arun Anandasivam
Priscilla Avegliano
Tina Balke
Sara Casare
Paul Chapron
Jochen Haller
Chung-Wei Hang
Chris Hazard
Stefan König
Guillaume Muller
Isaac Pinyol
Jochen Stoesser

Contents

1 Invited Papers

1.1 The Many Faces of Reputation: Towards a Science of Reputation System Design

Chris Dellarocas (Robert H. Smith School of Business)

Reputation systems are the unsung hero of the Web. They are the quiet force behind some of the most successful Internet companies: Reputation systems have driven the entire business model of eBay, are responsible to a large extent for Amazon's success and are fueling participation and loyalty in sites ranging from online communities to videogames. Their objectives are diverse, ranging from building trust to improving customer experience to motivating participation and effort. Despite all that, research on these systems has been highly fragmented and has focused on isolated questions, such as eliciting truthful opinions. There have been very few attempts to develop a comprehensive "big picture" framework on the diverse roles that reputation systems can play in Web x.0 initiatives, their building blocks and how they should be designed. My talk will survey work in this area and will offer the beginnings of such a framework.

1.2 Reputation in Evolution

Conte, Rosaria (ISTC-CNR)

Reputation will here be argued to have played a crucial role in the evolution of our species, thanks to its special social cognitive properties. After a brief report on the author's previous work in which reputation was defined as a meta-representation (Conte and Paolucci, 2002), some crucial properties of reputation will be derived, i.e. fast and cheap transmissibility. Next, reputation will be compared with other mechanisms of social control, i.e. strong reciprocity. Some hypotheses will be discussed, in particular (a) the higher tolerance of reputation with regard to liars (informational cheaters), which make more inclusive reputation-based groups, and (b) the more severe discrimination it allows of real or presumed material cheaters, which make a reputation unfalsifiable. The question as to which consequence this property bears on reputation-based groups will be discussed.

2 Foundations of Reputation

Information-Based Reputation

Carles Sierra[1] and John Debenham[2]

[1] Institut d'Investigacio en Intel.ligencia Artificial, Spanish Scientific Research Council, UAB
08193 Bellaterra, Catalonia, Spain sierra@iiia.csic.es
[2] University of Technology, Sydney, Australia debenham@it.uts.edu.au

Abstract. Information-based agents use tools from information theory to evaluate their utterances and to build their world model. When embedded in a social network these agents measure the strength of information flow in this sense. This leads to a model of information-based reputation in which agents share opinions, and observe the way in which their opinions effect the opinions of others. A method is proposed that supports the deliberative process of combining opinions into a group's reputation. The reliability of agents as opinion givers are measured in terms of the extent to which their opinions differ from that of the group reputation. These reliability measures are used to form an a priori reputation estimate given the individual opinions of a set of independent agents.

1 Introduction

Reputation measures are becoming a cornerstone of many applications over the web [1]. This is the case in recommender systems or in trading mediation sites. In these applications there is a need to assess, for instance, how much should we trust the recommendation coming from an unknown source, or how reliable a trading partner is. When the probability of having had previous interactions between two autonomous entities (agents), human or software, is very low, the use of reputation measures, i.e. group opinions, becomes a natural solution. In this paper we propose a reputation model that is inspired by information theory and that is based on the *information-based agency* explained elsewhere [2]. It also uses semantic distance over a shared ontology as a way to profit from similar experiences in the assessment of reputation, and we start to explore some social network analysis techniques to weigh the opinions of other agents according to their social relationships.

Reputation is the opinion (more technically, a social evaluation) of a group about something. So a group's reputation about a thing will be related in some way to the opinions that the individual group members hold towards that thing. An opinion is an assessment, judgement or evaluation of something, and are represented in this paper as probability distributions on a suitable ontology called the *evaluation space* E.

An opinion is an evaluation of an *aspect* of a thing. A rainy day may be evaluated as being "bad" from the aspect of being suitable for a picnic, and "good" from the aspect of watering the plants in the garden. An aspect is the "point of view" that an agent has when forming his opinion.

An opinion is evaluated in context. The *context* is the set of all things that the thing is being, explicitly or implicitly, evaluated with or against. The set of valuations of

all things in the context calibrates the valuation space. For example, "this is the best paper in the conference". The context can be vague: "of all the presents you could have given me, this is the best". If agents are to discuss opinions then they must have some understanding of each other's context.

Summarising the above, an *opinion* is an agent's evaluation of a particular aspect of a thing in context. A representation of an opinion will contain: the thing, its aspect, its context, and a distribution on E representing the evaluation of the thing.

In this paper we explore the case of opinions being formed through a social evaluation process illustrated in Figure 1. Each agent in a group of agents first forms an individual opinion on some thing. Second these individual opinions are shared with rest of the group. A group discussion follows as a result of which each agent states a revised opinion. Following that there is another discussion during which the group attempts to formulate a shared reputation for the thing. The model that we describe is based on three observations only for each participating agent: their initial individual opinion, their revised opinion, and the group's reputation if one is agreed upon. This social evaluation process was suggested by a process used to evaluate submissions to conferences.

2 The multiagent system

We assume that a multiagent system $\{\alpha, \beta_1, \ldots, \beta_o, \xi, \theta_1, \ldots, \theta_t\}$, contains an agent α that interacts with negotiating agents, β_i, information providing agents, θ_j, and an *institutional agent*, ξ, that represents the institution where we assume the interactions happen [3]. Institutions give a normative context to interactions that simplify matters (e.g an agent can't make an offer, have it accepted, and then renege on it). The institutional agent ξ may form opinions on the actors and activities in the institution and may publish reputation estimates on behalf of the institution. The agent ξ also fulfils a vital role to compensate for any lack of sensory ability in the other agents by promptly and accurately reporting observations as events occur; for an example, without such reporting an agent may have no way of knowing whether it is a fine day or not. When we consider the system from the point of view of a particular agent we will use agent α, and that is α's only significance.

Our agents are information-based [4], everything in their world is uncertain. To deal with this uncertainty, the world model, $\mathcal{M}^t$, consists of random variables each representing a point of interest in the world. Distributions are then derived for these variables on the basis of information received. Additionally, information-based agents [4] are endowed with machinery for valuing the information that they have, and that they receive. They were inspired by the observation that "everything an agent says gives away information". They model how much they know about other agents, and how much they believe other agents know about them. By classifying private information into functional classes, and by drawing on the structure of the ontology, they develop a map of the 'intimacy' [5] of their relationships with other agents.

2.1 Communication Model

We assume that all agents share an ontology O, that for simplicity we will consider as a set of well-formed expressions representing a given domain of discourse.[3]

An ontology is a tuple $O = (C, R, \leq, \sigma)$ where:

1. C is a finite set of concept symbols (including basic data types);
2. R is a finite set of relation symbols;
3. $\leq$ is a reflexive, transitive and anti-symmetric relation on C (a partial order)
4. $\sigma : R \rightarrow C^+$ is the function assigning to each relation symbol its arity

where $\leq$ is a traditional *is-a* hierarchy, and R contains relations between the concepts in the hierarchy.

Based on this ontology we define a simple language $\mathcal{C}$ that accounts for the expressions exchanged in gossiping dialogues, and is based on two fundamental primitives: experience$(\alpha, \beta, \varphi, \varphi')$ to represent, in φ, the world that β committed at bringing about and in φ' what α actually observed, and opinion$(\alpha, \beta, \varphi, o)$ to represent an opinion o that α makes about the behaviour or position of β with respect to φ. The opinion is expressed as a probability distribution p_i over a set of qualitative terms e_j. Experiences can also be considered argumentative moves in support of a particular opinion. Language $\mathcal{C}$ is then the set of utterances u defined as:

$$
\begin{aligned}
u &::= \text{inform}(agent, agent, content, time) \\
content &::= \text{opinion}(agent, agent, [term,](eval)) \mid \text{experience}(agent, agent, term, term) \\
term &::= \varphi|\phi| \ldots (*\textit{expression from ontology } O*) \\
eval &::= e = p \mid e = p, eval \\
e &::= good \mid bad \mid \ldots (*\textit{qualitative term}*) \\
p &::= a\ point\ in\ [0,1] \\
time &::= a\ point\ in\ time \\
agent &::= \alpha \mid \beta \mid \ldots (*\textit{agent identifiers}*)
\end{aligned}
$$

We will note by by $\mathcal{A}^t$ the set of existing agents at instant t, by $\mathcal{E}$ the set of all possible evaluation values e, and by Φ the set of all ontology-complaint terms. For example:

$\text{inform}(John, me, \text{opinion}(John, Carles, wrapping(package), (ghastly = 0.7)), t)$
$\text{inform}(John, me, \text{opinion}(Carles, John, suggesting(wine(Margaret\ River)), (excellent = 0.9)), t)$
$\text{inform}(John, me, \text{experience}(John, Carles, package(date(Monday)), package(date(Friday)), t)$
$\text{inform}(John, me, \text{experience}(John, Carles, fly(elephant), \neg fly(elephant)), t)$

The concepts within an ontology are closer, semantically speaking, depending on how far away are they in the structure defined by the $\leq$ relation. Semantic distance plays

[3] Local ontologies could also be considered together with appropriate ontology alignment techniques [6].

a fundamental role in strategies for information-based agency. How stated opinions, opinion$(\cdot)$, about objects in a particular semantic region, and their subsequent validation *affect* our decision making process about the significance of future opinions on nearby semantic regions is crucial to model the common sense that human beings apply. A measure [7] bases the *semantic similarity* between two concepts on the path length induced by $\leq$ (more distance in the $\leq$ graph means less semantic similarity), and the *depth* of the subsumer concept (common ancestor) in the shortest path between the two concepts (the deeper in the hierarchy, the closer the meaning of the concepts). Semantic similarity could then be defined as:

$$\mathrm{Sim}(c,c') = e^{-\kappa_1 l} \cdot \frac{e^{\kappa_2 h} - e^{-\kappa_2 h}}{e^{\kappa_2 h} + e^{-\kappa_2 h}}$$

where l is the length (i.e. number of hops) of the shortest path between the concepts, h is the depth of the deepest concept subsuming both concepts, and κ_1 and κ_2 are parameters scaling the contribution of shortest path length and depth respectively.

The following does not depend on this particular definition. Suppose an ontology is populated with probability distributions at each branch representing the preference in some sense of an agent; e.g. if wine $\geq$ {red wine, white wine} then the probability at that branch could represent Carles' preference for red or white wine. If the same ontology is populate with John's probabilities then a metric such as the Kullback-Leibler [8] divergence can be used to measure the difference in the significance of the term 'red wine' to Carles and to John.

2.2 The Social Structure of the Multiagent System

Agents, or groups, in an evolving network can be described by a number of measures of their importance or prominence [9] [10]. These measures summarise the structural relations among all nodes in the network and account for an agent's choices (whom do I link to) as well as the other agent's choices (who links to me). Centrality measures try to determine prominence by not taking into account the direction of the ties, and prestige measures when direction matters. Given a matrix $R(n,n)$ that represents in $r_{ij} \in [0,1]$ the intensity of the relation R from i to j we define:

- Centrality measures. Determining in how many relationships a particular agent is involved.
  - Normalised Degree Centrality. The extent to which a node connects to the rest.

$$C_d(i) = \frac{\sum_{j=1}^{n} r_{ij}}{n-1}$$

  - Normalised Closeness Centrality. How near a node is from the rest.

$$C_c(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}$$

    where $d(i,j)$ is the minimum distance between i and j in the graph

- Normalised Betweenness Centrality. The extent to which an agent lies on the shortest paths between pairs of agents in the graph.

$$C_b(i) = \frac{2}{(n-1)(n-2)} \cdot \sum_{j,k \neq i, j \neq k} \frac{s_{jk}(i)}{s_{jk}}$$

  where $s_{jk}(i)$ is the number of shortest paths between j and k including i, and s_{jk} is the total number of shortest paths between j and k.

– Prestige Degree. Determining how many links an agent receives.

$$P(i) = \frac{\sum_{j=1}^{n} r_{ji}}{n-1}$$

The preceding measures are topological and do not capture what the connections between the individual agents are used for. From the perspective of information-based agency we are interested in two things: first, how much information is passing along the connections, and second, the value of that information to the receiving agent.

The 'Source Coding Theorem' of Shannon states that N independent, identically-distributed random variable each with entropy $\mathbb{H}(X)$ can be compressed into marginally more than $N \times \mathbb{H}(X)$ bits of information. In other words, if we know the amount of information that has been transmitted in bits, and that the coding is loss-less then we know the amount of information that has been transmitted in terms of the lack of uncertainty that it could bring. Further, if we have mutually exclusive events, E_i, each with prior probabilities, p_i, then the *expected information content* $\mathbb{I}$ of a message that transforms the priors p_i into posterior probabilities q_i is: $\mathbb{I} = \sum_i q_i \times \log(\frac{q_i}{p_i})$. These ideas enable use to analyse network structure from the perspective of information flow. [11] defines the *path-transfer centrality* of vertex i as $-\sum_j p_{ij} \log p_{ij}$ where p_{ij} is the probability that a communication path starting at node i will end at node j. If an agent receives a message containing information $\mathbb{I}$ then the Shannon *value* of $\mathbb{I}$ is: $\mathbb{H}(\mathcal{M}^t \,|\, I) - \mathbb{H}(\mathcal{M}^t)$, where $\mathcal{M}^t$ is the agent's world model. When used together with the ontology and a map of $\mathcal{M}^t$ that categorises the agent's information, this measure can be used to take stock of the information in $\mathcal{M}^t$.

3 Forming Opinions

This section describes how an information-based agent forms opinions. Section 4 will describe how the opinions of the agents in a group may be distilled into a reputation.

An opinion is a valuation by an agent of an aspect of a thing taken in context. Formally, $O_i(z,a,C)$ represents the result of the valuation by agent β_i of aspect a of thing z in context C. For example, the valuation by agent "Carles" of the "scientific quality" aspect of the thing "John's paper" in the context of "the AAMAS conference submissions". Opinions are communicated using the language described in Section 2.1. The context C is often subjectively chosen by the agent, and is not part of the opinion$(\cdot)$ primitive, although context may be the subject of associated argumentation. For example, re-using an example of communication from Section 2.1:

$$\text{inform}(John, me, \text{opinion}(John, Carles, wrapping(package), (ghastly = 0.7)), t)$$

we can extract an opinion as:

$$O(package, wrapping, the\ way\ I\ do\ wrapping) = (ghastly = 0.7)$$

As noted above, to preserve consistency and generality we assume that all opinions are expressed as probability distributions over some suitable E. If an agent expresses an opinion as $\mathbb{P}(X = e_i)$ we treat this as the distribution with minimum relative entropy with respect to the prior subject to the constraint $\mathbb{P}(X = e_i)$ — in case there is no known prior we use the maximum entropy, uniform distribution. For example, if $E = (\text{fine}, \text{cloudy}, \text{wet}, \text{storm})$ then the opinion "I am 70% certain that tomorrow will be fine" will be represented as $(0.7, 0.1, 0.1, 0.1)$ for a uniform prior.

The distributions in an agent's world model $\mathcal{M}^t$ represent the agent's opinions about the value of the corresponding random variable over some valuation space. Opinions may be derived from opinions. For example, to form an opinion on "tomorrow's suitability for a picnic" and agent may introduce random variables for: tomorrow's mid-day temperature, tomorrow's mid-day cloud cover, and tomorrow's mid-day wind strength, construct distributions for them using on-the-fly weather forecast information, and then derive an opinion about the picnic somehow from these three distributions.

In Section 3.1 we describe how the distributions in the world model are updated as real-time information becomes available; in that section we also estimate the reliability of each information source by subsequently validating the information received from it.

3.1 Updating Opinions with Real-Time Information

In the absence of in-coming messages the distributions in $\mathcal{M}^t$ should gradually decay towards some zero-information state. In many cases there is background knowledge about the world — for example, a distribution of the daily maximum temperature in Barcelona in May — such a distribution is called a *decay-limit distribution*. If the background knowledge is incomplete then one possibility is to assume that the decay limit distribution has maximum entropy whilst being consistent with the available data. Given a distribution, $\mathbb{P}(X_i)$, and a decay limit distribution $\mathbb{D}(X_i)$, $\mathbb{P}(X_i)$ decays by:

$$\mathbb{P}^{t+1}(X_i) = \Delta_i(\mathbb{D}(X_i), \mathbb{P}^t(X_i)) \tag{1}$$

where Δ_i is the *decay function* for the X_i satisfying the property that $\lim_{t\to\infty} \mathbb{P}^t(X_i) = \mathbb{D}(X_i)$. For example, Δ_i could be linear: $\mathbb{P}^{t+1}(X_i) = (1-\nu_i) \times \mathbb{D}(X_i) + \nu_i \times \mathbb{P}^t(X_i)$, where $\nu_i < 1$ is the decay rate for the i'th distribution. Either the decay function or the decay limit distribution could also be a function of time: Δ_i^t and $\mathbb{D}^t(X_i)$.

The following procedure updates $\mathcal{M}^t$. Suppose that α receives a message μ from agent β at time t.[4] Suppose that this message states that something is so with probability v, and suppose that α attaches an epistemic belief $\mathbb{R}^t(\alpha, \beta, \mu)$ to μ — this probability reflects α's level of personal *caution*. Each of α's active plans, s, contains constructors for a set of distributions $\{X_i\} \in \mathcal{M}^t$ together with associated *update functions*, $J_s(\cdot)$,

[4] This message is not necessarily a message from the language in section 2.1. We refer with μ to any *inform* message with propositional content that can be processed by the agent.

such that $J_s^{X_i}(\mu)$ is a set of linear constraints on the posterior distribution for X_i. Denote the prior distribution $\mathbb{P}^t(X_i)$ by p, and let $p_{(\mu)}$ be the distribution with minimum relative entropy[5] with respect to p: $p_{(\mu)} = \arg\min_r \sum_j r_j \log \frac{r_j}{p_j}$ that satisfies the constraints $J_s^{X_i}(\mu)$. Then let $q_{(\mu)}$ be the distribution:

$$q_{(\mu)} = \mathbb{R}^t(\alpha,\beta,\mu) \times p_{(\mu)} + (1 - \mathbb{R}^t(\alpha,\beta,\mu)) \times p \tag{2}$$

and then let:

$$\mathbb{P}^t(X_{i(\mu)}) = \begin{cases} q_{(\mu)} & \text{if } q_{(\mu)} \text{ is more interesting than } p \\ p & \text{otherwise} \end{cases} \tag{3}$$

A general measure of whether $q_{(\mu)}$ is more interesting than p is: $\mathbb{K}(q_{(\mu)} \| \mathbb{D}(X_i)) > \mathbb{K}(p \| \mathbb{D}(X_i))$, where $\mathbb{K}(x\|y) = \sum_j x_j \ln \frac{x_j}{y_j}$ is the Kullback-Leibler divergence between two probability distributions x and y.

Finally merging Eqn. 3 and Eqn. 1 we obtain the method for updating a distribution X_i on receipt of a message μ:

$$\mathbb{P}^{t+1}(X_i) = \Delta_i(\mathbb{D}(X_i), \mathbb{P}^t(X_{i(\mu)})) \tag{4}$$

This procedure deals with integrity decay, and with two probabilities: first, the probability v in the message μ, and second the belief $\mathbb{R}^t(\alpha,\beta,\mu)$ that α attached to μ.

Reliability of the Information Source. An empirical estimate of $\mathbb{R}^t(\alpha,\beta,\mu)$ may be obtained by measuring the 'difference' between commitment and verification. Suppose that μ is received from agent β at time u and is verified by ξ as μ' at some later time t.[6] Denote the prior $\mathbb{P}^u(X_i)$ by p. Let $p_{(\mu)}$ be the posterior minimum relative entropy distribution subject to the constraints $J_s^{X_i}(\mu)$, and let $p_{(\mu')}$ be that distribution subject to $J_s^{X_i}(\mu')$. We now estimate what $\mathbb{R}^u(\alpha,\beta,\mu)$ should have been in the light of knowing *now*, at time t, that μ should have been μ'.

The idea of Eqn. 2, is that $\mathbb{R}^t(\alpha,\beta,\mu)$ should be such that, *on average* across $\mathcal{M}^t$, $q_{(\mu)}$ will predict $p_{(\mu')}$ — no matter whether or not μ was used to update the distribution for X_i, as determined by the condition in Eqn. 3 at time u. The *observed reliability* for μ and distribution X_i, $\mathbb{R}^t_{X_i}(\alpha,\beta,\mu)|\mu'$, on the basis of the verification of μ with μ', is the value of k that minimises the Kullback-Leibler divergence:

$$\mathbb{R}^t_{X_i}(\alpha,\beta,\mu)|\mu' = \arg\min_k \mathbb{K}(k \cdot p_{(\mu)} + (1-k) \cdot p \parallel p_{(\mu')})$$

[5] Given a probability distribution q, the *minimum relative entropy distribution* $p = (p_1,\ldots,p_I)$ subject to a set of J linear constraints $g = \{g_j(p) = a_j \cdot p - c_j = 0\}, j = 1,\ldots,J$ (that must include the constraint $\sum_i p_i - 1 = 0$) is: $p = \arg\min_r \sum_j r_j \log \frac{r_j}{q_j}$. This may be calculated by introducing Lagrange multipliers λ: $L(p,\lambda) = \sum_j p_j \log \frac{p_j}{q_j} + \lambda \cdot g$. Minimising L, $\{\frac{\partial L}{\partial \lambda_j} = g_j(p) = 0\}, j = 1,\ldots,J$ is the set of given constraints g, and a solution to $\frac{\partial L}{\partial p_i} = 0, i = 1,\ldots,I$ leads eventually to p. Entropy-based inference is a form of Bayesian inference that is convenient when the data is sparse [12] and encapsulates common-sense reasoning [13].

[6] This could be later communicated as inform$(\gamma,\alpha,$experience$(\gamma,\beta,\mu,\mu'),t)$.

The predicted *information* in the enactment of μ with respect to X_i is:

$$\mathbb{I}^t_{X_i}(\alpha,\beta,\mu) = \mathbb{H}^t(X_i) - \mathbb{H}^t(X_{i(\mu)}) \tag{5}$$

that is the reduction in uncertainty in X_i where $\mathbb{H}(\cdot)$ is Shannon entropy. Eqn. 5 takes account of the value of $\mathbb{R}^t(\alpha,\beta,\mu)$.

If $\mathbf{X}(\mu)$ is the set of distributions that μ affects, then the *observed reliability* of β on the basis of the verification of μ with μ' is:

$$\mathbb{R}^t(\alpha,\beta,\mu)|\mu' = \frac{1}{|\mathbf{X}(\mu)|}\sum_i \mathbb{R}^t_{X_i}(\alpha,\beta,\mu)|\mu' \tag{6}$$

If $\mathbf{X}(\mu)$ are independent the predicted *information* in μ is:

$$\mathbb{I}^t(\alpha,\beta,\mu) = \sum_{X_i \in \mathbf{X}(\mu)} \mathbb{I}^t_{X_i}(\alpha,\beta,\mu) \tag{7}$$

Suppose α sends message μ to β where μ is α's private information, then assuming that β's reasoning apparatus mirrors α's, α can estimate $\mathbb{I}^t(\beta,\alpha,\mu)$.

For each formula φ at time t when μ has been verified with μ', the *observed reliability* that α has for agent β in φ is:

$$\mathbb{R}^{t+1}(\alpha,\beta,\varphi) = (1-\nu)\times\mathbb{R}^t(\alpha,\beta,\varphi) + \nu\times\mathbb{R}^t(\alpha,\beta,\mu)|\mu'\times \mathrm{Sim}(\varphi,\mu)$$

where Sim measures the semantic distance between two sections of the ontology as introduced in Section 2.1, and ν is the learning rate. Over time, α notes the context of the various μ received from β, and over the various contexts calculates the relative frequency, $\mathbb{P}^t(\mu)$. This leads to an overall expectation of the *reliability* that agent α has for agent β:

$$\mathbb{R}^t(\alpha,\beta) = \sum_\mu \mathbb{P}^t(\mu)\times\mathbb{R}^t(\alpha,\beta,\mu)$$

3.2 Verifiable Opinions

An opinion is *verifiable* if within a "reasonable amount of time" it ceases to be an opinion and becomes an observable fact; for example, the opinion "tomorrow's maximum temperature will be over 30°" is verifiable, whereas the opinion "the Earth will exist in 100,000 years time" is not verifiable in any practical sense, and "Brahms' symphonies are ghastly" will never be verifiable.

The articulation by β of a verifiable opinion carries with it the intrinsic commitment that it will in due time become an observable true fact. α will be interested in any variation between β's commitment, φ, and what is actually observed (as advised by the institution agent ξ), as the fact, φ'. We denote the relationship between opinion and fact, $\mathbb{P}^t(\mathrm{Observe}(\varphi')|\mathrm{Commit}(\varphi))$ simply as $\mathbb{P}^t(\varphi'|\varphi) \in \mathcal{M}^t$.

In the absence of in-coming messages the conditional probabilities, $\mathbb{P}^t(\varphi'|\varphi)$, should tend to ignorance as represented by the *decay limit distribution* and Eqn. 1. We now show how Eqn. 4 may be used to revise $\mathbb{P}^t(\varphi'|\varphi)$ as observations are made. Let the

set of possible factual outcomes be $\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_m\}$ with prior distribution $p = \mathbb{P}^t(\varphi'|\varphi)$. Suppose that message μ is received from ξ that verifies or refutes a previously stated verifiable opinion expressed by β, we estimate the posterior $p_{(\mu)} = (p_{(\mu)i})_{i=1}^m = \mathbb{P}^{t+1}(\varphi'|\varphi)$.

First, if $\mu = (\varphi_k, \varphi)$ is observed then α may use this observation to estimate $p_{(\varphi_k)k}$ as some value d at time $t+1$. We estimate the distribution $p_{(\varphi_k)}$ by applying the principle of minimum relative entropy as in Eqn. 4 with prior p, and the posterior $p_{(\varphi_k)} = (p_{(\varphi_k)j})_{j=1}^m$ satisfying the single constraint: $J^{(\varphi'|\varphi)}(\varphi_k) = \{p_{(\varphi_k)k} = d\}$.

Second, we consider the effect that the verification ϕ' of another simple, verifiable opinion ϕ of β has on p. This is achieved by appealing to the structure of the ontology using the $\mathrm{Sim}(\cdot)$ function. Given the observation $\mu = (\phi', \phi)$, define the vector t by:

$$t_i = \mathbb{P}^t(\varphi_i|\varphi) + (1 - \mid \mathrm{Sim}(\phi', \phi) - \mathrm{Sim}(\varphi_i, \varphi) \mid) \cdot \mathrm{Sim}(\varphi', \phi)$$

for $i = 1, \ldots, m$. t is not a probability distribution. The multiplying factor $\mathrm{Sim}(\varphi', \phi)$ limits the variation of probability to those formulae whose ontological context is not too far away from the observation. The posterior $p_{(\phi', \phi)}$ is defined to be the normalisation of t.

In this section we have shown how an information-based agent models the accuracy of an agent's opinions when they are verifiable. The model produced is predictive in the sense that when an opinion is stated it gives a distribution of expectation over the space of factual outcomes.

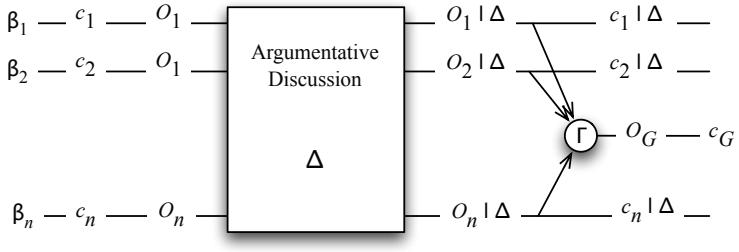
3.3 Unverifiable Opinions

If an opinion can not be verified then one way in which it may be evaluated is to compare it with the corresponding individual opinions, or group reputation, of a group of agents. The focus of this paper is on reputation; that is, a social evaluation conducted by a group. We deal with unverifiable opinions using a social evaluation framework that is abstracted from any particular case and is illustrated in Figure 1. The idea is that a group G of n agents independently form a prior opinion, O_i on the same thing. Each agent has a prior confidence value, c_i, that estimates how close its prior opinion, O_i, is expected to be to the reputation, or common opinion, of the group, R_G — precisely c_i measures how effective the agent is at influencing the opinions of other agents, it does *not* measure how good its opinion is in any absolute sense as the opinion is assumed to be unverifiable. The agents then make their prior opinions public to the other agents and an argumentative discussion, Δ, takes place during which the agents may choose to revise their opinions, $O_i|\Delta$. When the revised opinions are published a second argumentative discussion, Γ, takes place during which the agents attempt to distil their opinions into a group reputation, R_G. The confidence estimates, c_i are then revised by noting the differences between O_i, $O_i|\Delta$ and R_G, to give posterior values, $c_i|\Delta$. The processes in Figure 1 are summarised as:

$$\Delta : f(\{(O_i, c_i\}) = \{O_i|\Delta\}$$

$$\Gamma : g(\{(O_i|\Delta, c_i\}) = (R_G, d_G)$$

Fig. 1. The social evaluation framework in which a group G of n agents $\beta_1,\ldots,\beta_n$ table their private opinions $O_1,\ldots,O_n$, have an open, argumentative discussion Δ (see Section 3.3), and then revise their opinions $O_1|\Delta,\ldots,O_n|\Delta$. This is followed by another argumentative discussion Γ (see Section 4) during which the agents consider whether revised opinions can be distilled into a common reputation R_G. The symbols c_i and c_G are confidence values as explained below.



$$\{\Delta,\Gamma\} : h(\{(O_i,c_i,O_i|\Delta\},R_G) = \{c_i|\Delta\}$$

The function $f(\cdot)$ is the product of the discussion Δ — we simply observe the outcome. Function $g(\cdot)$ is described in Section 4, and $h(\cdot)$ in Section 5.

4 Combining Opinions and Forming Reputation

A reputation is a social evaluation by a group. When the group is a set of autonomous agents the only sense in which an opinion can exist is as a common opinion throughout the group. The objective of the argumentative process Γ in Figure 1 is to determine a common view if one exists. The following procedure first determines whether a common view exists, and second it offers three views of what that common view could be. The three different views vary with differing degrees of statistical dependence between the agents.

The process of distilling opinions into a reputation can not simply be computed. For example, consider two agents who are reviewing the same conference paper and are in total agreement about the result "a 'strong accept' with confidence 0.8" where the reliability of each agent is 90%. What should their combined opinion, or in this case 'paper reputation', be? As their individual reliability is 90% perhaps the common view is "a 'strong accept' with confidence 0.72". Alternatively because they both agree, and may have quite different reasons supporting their views, perhaps the common view should be "a 'strong accept' with some confidence greater than 0.8".

The work described in the remainder of this section and in Section 5 is expressed in terms of two agents; it extends naturally to n agents. The procedure is based on three methods that are detailed below.

Dependent Method. To form a *combined opinion* of two opinions, X_1 and X_2, construct the joint distribution $W = (X_1, X_2, Z)$ and impose the constraints:

$$\left(\sum_i \mathbb{P}(W = w_i) \mid X_k = x_j\right) = \mathbb{P}(X_k = x_j),\ k = 1,2$$

$$\left(\sum_i \mathbb{P}(W = w_i) \mid X_k = Z\right) = c_k,\ k = 1,2$$

let W be the distribution of maximum entropy that satisfies these constraints. Then the combined opinion $\mathrm{Dep}(X_1, X_2)$ is $\mathbb{P}(Z = z)$. If the data is inconsistent then the value is undefined — this is a test of whether the data is consistent. If the data is inconsistent then this indicates that there is no shared opinion. Being based on a maximum entropy calculation the posterior is a conservative combination of the given opinions — it is "maximally noncommittal" to that which is not known. To calculate this dependent, combined opinion when the prior is known, calculate the minimum relative entropy distribution with respect to that prior using the same constraints as described.

Υ *Method.* Let's define $\mathbb{P}(\alpha, d)$ as the probability that an opinion O_α expressed by α (i.e, a probability distribution) is at distance d of the true distribution (or at distance d of a group opinion). That is, the probability that a certain distribution Q is the right one is defined as $P(Q \text{ is right}) = \mathbb{P}(\alpha, DIST(O_\alpha, Q))$ for an appropriate distance measure $DIST$.[7] These distributions can be obtained by datamining past group opinion formation processes.

Given a group G, we look for the group opinion, R_G such that the certainty on that group opinion being the right one is maximised. That is,

$$R_G = \max_Q \Upsilon(\{\mathbb{P}(\alpha, DIST(O_\alpha, Q))\}_{\alpha \in G})$$

Where Υ is the uninorm operator [15]. In case there are several such group opinions we prefer the one with maximum entropy. And then,

$$d_G = \Upsilon(\{\mathbb{P}(\alpha, DIST(O_\alpha, R_G))\}_{\alpha \in G})$$

For the values in Table 1, we discreetise the $\mathbb{P}(\alpha, d)$ in the intervals between the points in the following list: $[0, 0.035, 0.3, 0.5, 0.8, 1]$.

Independent Method. Given a prior distribution $\mathbb{P}(W = x_j)$, a pair of opinions, $\mathbb{P}(X_i = x_j)$ $i = 1,2$, with their respective certainties c_i, assuming that the agents' opinions are statistically independent, let $w_{i,j} = c_i \times \mathbb{P}(X_i = x_j)$, $i = 1,2$, and let $v_j = \frac{\prod_i w_{i,j}}{\prod_i w_{i,j} + \prod_i (1 - w_{i,j})}$ then the combined opinion $\mathrm{Ind}(X_1, X_2)$ is: $v_j + (1 - \sum_k v_k) \times \mathbb{P}(W = x_j)$, with strength $\sum_k v_k$. This method assumes that the priors are independent (unlikely in practice) and has the property that the probabilities in two similar distributions are amplified.

The overall procedure plays the role of a mediator. If the 'Dependent Method' does not return a value then the data is inconsistent, and the agents should either have further

[7] Kullback-Leibler divergence, or the earth movers distance [14] could be used.

discussion or "agree to disagree". Otherwise calculate the three values $\mathrm{Dep}(\cdot)$, $\Upsilon(\cdot)$ and $\mathrm{Ind}(\cdot)$. Propose $\Upsilon(\cdot)$ to the agents, and if they accept it then that is their common opinion. Otherwise propose that their common opinion lies somewhere between $\mathrm{Dep}(\cdot)$ and $\mathrm{Ind}(\cdot)$ and leave it to them to determine it.

Table 1 contains some sample values for the three methods. In Case 3 the two opinions are identical with maximal value of 0.8 and strengths of 0.8 and 0.9. The $\mathrm{Dep}(X_1, X_2)$ method is conservative and gives 0.77 because of the strength values. The $\Upsilon(X_1, X_2)$ method balances the strength uncertainty with the fact that their are two shared views to give 0.8. The $\mathrm{Ind}(X_1, X_2)$ method is bold and gives 0.85 because two agents share the same view; the boldness of the $\mathrm{Ind}(X_1, X_2)$ method is balanced by its comparatively low strength values.

Table 1. Three cases of sample values for the three methods for combining opinions. In each case the opinions are X_1 and X_2 and the strength of the distributions is denoted by "Str". The right hand column contains the discreetised $\mathbb{P}(\alpha, d)$ values described in the 'Υ Method'. All calculations were performed with a uniform prior.

Case 1	X_1	0.1000	0.5000	0.2000	0.1000	0.1000	Str = 0.9	$P = \langle 0.9, 0.05, 0.03, 0.01, 0.01 \rangle$
	X_2	0.0500	0.8000	0.0500	0.0500	0.0500	Str = 0.7	$P = \langle 0.7, 0.2, 0.05, 0.03, 0.02 \rangle$
	Dep	0.0919	0.5590	0.1653	0.0919	0.0919	$c_G \approx 1$	
	Υ	0.0700	0.7000	0.1700	0.0700	0.0700	$c_G = 0.95$	
	Ind	0.0978	0.6044	0.1022	0.0978	0.0978	$c_G = 0.53$	
Case 2	X_1	0.1000	0.6000	0.1000	0.1000	0.1000	Str = 0.8	$P = \langle 0.8, 0.1, 0.04, 0.01, 0.01 \rangle$
	X_2	0.0500	0.8000	0.0500	0.0500	0.0500	Str = 0.9	$P = \langle 0.9, 0.06, 0.03, 0.01, 0.01 \rangle$
	Dep	0.0683	0.7266	0.0683	0.0683	0.0683	$c_G \approx 1$	
	Υ	0.08	0.63	0.08	0.08	0.08	$c_G = 0.97$	
	Ind	0.0601	0.7596	0.0601	0.0601	0.0601	$c_G = 0.72$	
Case 3	X_1	0.0500	0.8000	0.0500	0.0500	0.0500	Str = 0.8	$P = \langle 0.8, 0.1, 0.04, 0.01, 0.01 \rangle$
	X_2	0.0500	0.8000	0.0500	0.0500	0.0500	Str = 0.9	$P = \langle 0.9, 0.06, 0.03, 0.01, 0.01 \rangle$
	Dep	0.0573	0.7707	0.0573	0.0573	0.0573	$c_G \approx 1$	
	Υ	0.05	0.8	0.05	0.05	0.05	$c_G = 0.97$	
	Ind	0.0363	0.8548	0.0363	0.0363	0.0363	$c_G = 0.83$	

5 Reputation of the Agents

In the previous section we described how a mediator could assist agents to agree on a common opinion, or reputation, of some thing being evaluated. Additionally, the institution ξ builds a view of the reputation of the individual agents who perform the evaluations by observing the process illustrated in Figure 1. In particular, ξ observes the development of the c_i values (described below), the distances between initial opinion O_i and considered opinion $O_i|\Delta$, and the distances between both opinions and the group reputation R_G when it exists.

Given two opinions X_1 and X_2 the *strength* of X_1 on X_2 is defined as: $\mathbb{P}(X_1 = X_2)$. If X_1 and X_2 are both defined over the same valuation space $E = \{e_i\}_{i=1}^{n}$ then: $\mathbb{P}(X_1 =$

$X_2) = \sum_i P(W = w_i) \mid X_1 = X_2$, where $W = (X_1, X_2)$ is the joint distribution. That is, we sum along the diagonal of the joint distribution. We estimate the diagonal w_i values using the dependent estimate: $\mathbb{P}(X_1 = e_i) \wedge \mathbb{P}(X_1 = e_i) = \min_j \mathbb{P}(X_j = e_i)$, and hence: $\text{Str}(X_1, X_2) = \mathbb{P}(X_1 = X_2) = \sum_i \min_j \mathbb{P}(X_j = e_i)$. A measure of the distance between X_1 and X_2 is then: $\text{Dist}(X_1, X_2) = 1 - \text{Str}(X_1, X_2)$. This definition of strength is consistent with the 'Dependent Method' in Section 4 that is the basis of the reputation mediation procedure. Other definitions include the Kullback-Leibler divergence, $\text{Dist}(X_1, X_2) = \mathbb{K}(X_1 || X_2)$, and the earth movers distance [14].

Each time a reputation R_G is formed, the c_i values are updated using: $c_i|\Delta = \mu \times \text{Dist}(O_i, R_G) + (1 - \mu) \times c_i$, where μ is the learning rate. These c_i values are the product of successive social evaluation processes, and so they are reputation estimates.

The measures described above do not take the structure of the evaluation space E into account. Four additional measures are:

A generic distance measure. $\text{Dist}(X, Y) = \mathbb{K}(X' || Y')$ where (X', Y') is a permutation of (X, Y) the satisfies $X' < Y'$, and the order is defined by: $R_G < O_i|\Delta < O_i$. I.e. the earliest occurring distribution "goes in the second argument". This complication with ordering is necessary because $\mathbb{K}$ is not symmetric; it attempts to exploit the sense of relative entropy. An alternative is to use the symmetric form as it was originally proposed: $\frac{1}{2}(\mathbb{K}(X, Y) + \mathbb{K}(Y, X))$

A distance measure when the prior, Z, is known. This builds on the generic measure, and captures the idea that the distance between a pair of unexpected distributions is greater than the difference between a pair of similar, expected distributions. We measure of how expected X is by: $\mathbb{K}(X, Z)$, and normalise it by: $max_I \mathbb{K}(I, Z)$ to get: $e(X) = \frac{\mathbb{K}(X,Z)}{max_I \mathbb{K}(I,Z)}$. Then this measure is the arithmetic product of the previous generic measure with: $\frac{e(X)+e(Y)}{2}$.

A semantic distance measure. Suppose there is a difference measure $\text{Diff}(\cdot, \cdot)$ defined between concepts in the ontology — it could be related to the $\text{Sim}(\cdot, \cdot)$ function in Section 2.1. Then the distance between two opinions X and Y over valuation space E (represented as distributions p_i and q_i respectively) is: $\text{Dist}(X, Y) = \sum_{ij} p_i \times q_j \times \text{Diff}(e_i, e_j)$ where e_i are the categories in E.

A distance measure when E is ordered and the prior is known. If the valuation space E has a natural order, and if there is a known prior then define $\text{Diff}(e_i, e_j)$ to be the proportion of the population that is expected to lie between e_i and e_j. Then define $\text{Dist}(X, Y) = \sum_{ij} p_i \times q_j \times \text{Diff}(e_i, e_j)$. For example, in conference reviewing, if the expectation is that 40% of reviews are 'weak accept' and 20% are 'accept' then Diff('weak accept', 'accept')= $\frac{40}{2} + \frac{20}{2}$; i.e. taking the mid points of the intervals.

The measures described for $\text{Dist}(X, Y)$ are now used to enable ξ to attribute various reputations to agents. These reputation measures all assume that the agents have been involved in a number of successive social evaluation rounds as shown in Figure 1.

Inexorable. If agent β_i is such that: $\text{Dist}(O_i, O_i|\Delta) \ll \text{Dist}(O_i, O_j|\Delta), \forall j \neq i$ consistently holds then β_i is *inexorable*.

Predetermination. If: $\mathrm{Dist}(O_i, R_G) \ll \mathrm{Dist}(O_j, R_G), \forall j \neq i$ consistently, then β_i is a good '*predeterminer*'. Such an agent will have a high c_i value.

Persuasiveness. If β_i is such that: $\mathrm{Dist}(O_i, O_j|\Delta) \ll \mathrm{Dist}(O_j, O_j|\Delta), \forall j \neq i$ consistently then β_i is *persuasive*.

Compliance. If β_i is such that: $O_i|\Delta \approx \arg\min_X \sum_{j \neq i} \mathrm{Dist}(O_j|\Delta, X)$, then β_i is *compliant*.

Dogmatic. If β_i is such that: $O_i = O_i|\Delta$ consistently then β_i is *dogmatic*. A dogmatic agent is highly inexorable.

Adherence. If β_i is such that $O_i|\Delta = O_j$ where $j = \arg\max_{k,k \neq i} c_k$ consistently then β_i is *adherent* (in this round adherent to agent β_j).

6 Discussion

This paper has proposed a number of methods to ground the social building of reputation measures. The methods are based on information theory and permit to combine opinions when there is a high level of independence in the formation of the individual opinions. The method permits the computation of reputation values as aggregation of individual opinions, and also detects when agreement is not feasible. This impossibility may be used to trigger further discussions among the members of the group or to introduce changes in the composition of the group to permit agreements.

The use of social network analysis measures permits to define heuristics on how to combine opinions when there is no complete independence in the opinions expressed by the agents. There are a number of different relationships that may be used to guess dependency. For instance, in the context of scientific publications, co-authorship or affiliation, meaning that authors have written papers together or belong to the same laboratory may indicate a significant exchange of information between them and therefore a certain level of dependency. The aggregation of values by function h can then use these measures to diminish the joint influence of dependent opinions into the reputation. This is to be explored in future extensions of the information based reputation model.

Also, social networks can be used to assess initial values of c_i, the confidence on agent's opinions. For instance, we can say that an individual is expert in an area (keyword) if it is author of highly cited papers on the topic, has reviewed prestigious papers on the area, and has a central role in the college. This is easily expressed as

$$c_i = f\left(\sum_{\substack{(i,p)\in Authorship,\\(p,X)\in Area}} P^{Citation}(p), \sum_{\substack{(i,p)\in Review,\\(p,X)\in Area}} P^{Citation}(p), C_b^{College}(i)\right)$$

where $(i,p) \in Authorship$ means that agent i is author of paper p, $(p,X) \in Area$ means that paper p is on topic X and $(i,p) \in Review$ means that agent i has reviewed paper p. $Citation$ relates papers and $College$ relates authors. See Section 2.2 for definitions of P and C_b.

Our future work will include the in depth analysis of Social Network Measures in the information model reputation and the experimental analysis of the model in the context of scientific publishing as planned in the LiquidPub project (http://www.liquidpub.org).

Also, we will analyse the robustness of the proposed model in front of strategic reasoners that may try and manipulate the scores to their benefit.

References

1. Sabater, J., Sierra, C.: Review on computational trust and reputation models. Artificial Intelligence Review **24**(1) (September 2005) 33–60
2. Sierra, C., Debenham, J.: Information-based agency. In: Proceedings of Twentieth International Joint Conference on Artificial Intelligence IJCAI-07, Hyderabad, India (January 2007) 1513–1518
3. Arcos, J.L., Esteva, M., Noriega, P., Rodríguez, J.A., Sierra, C.: Environment engineering for multiagent systems. Journal on Engineering Applications of Artificial Intelligence **18** (2005)
4. Sierra, C., Debenham, J.: Information-based agency. In Veloso, M., ed.: Twentieth International Joint Conference on AI. (January 2007) In press
5. Sierra, C., Debenham, J.: The LOGIC Negotiation Model. In: Proceedings Sixth International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2007, Honolulu, Hawai'i (May 2007) 1026–1033
6. Kalfoglou, Y., Schorlemmer, M.: IF-Map: An ontology-mapping method based on information-flow theory. In Spaccapietra, S., March, S., Aberer, K., eds.: Journal on Data Semantics I. Volume 2800 of Lecture Notes in Computer Science. Springer-Verlag: Heidelberg, Germany (2003) 98–127
7. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering **15**(4) (July / August 2003) 871 – 882
8. MacKay, D.: Information Theory, Inference and Learning Algorithms. Cambridge University Press (2003)
9. Knoke, D., Yang, S.: Social Network Analysis. Sage Publications, Inc., California (2008)
10. Friedkin, N., Johnsen, E.: Social influence networks and opinion change. Advances in Group Processes **16** (1999) 1–29
11. Tutzauer, F.: Entropy as a measure of centrality in networks characterized by path-transfer flow. Social Networks **29**(2) (May 2007) 249–265
12. Cheeseman, P., Stutz, J.: On The Relationship between Bayesian and Maximum Entropy Inference. In: Bayesian Inference and Maximum Entropy Methods in Science and Engineering. American Institute of Physics, Melville, NY, USA (2004) 445 – 461
13. Paris, J.: Common sense and maximum entropy. Synthese **117**(1) (1999) 75 – 93
14. Peleg, S., Werman, M., Rom, H.: A unified approach to the change of resolution: Space and gray-level. IEEE Transactions on Pattern Analysis and Machine Intelligence **11**(7) (July 1989) 739—742
15. Yager, R.R.: On the determination of strength of belief for decision support under uncertainty — part i: generating strengths of belief. Fuzzy Sets and Systems **142**(1) (2004) 117–128

Cooperation and Reputation in Dynamic Networks *

Rense Corten[1] and Karen S. Cook[2]

[1] Department of Sociology/ICS, Utrecht University, r.corten@uu.nl
[2] Department of Sociology, Stanford University, k.cook@stanford.edu

Abstract. A common finding is that cooperation is more likely if interactions are embedded in cohesive social networks, which is often explained by reputation mechanisms. An underlying assumption in these explanations is that networks are exogenously imposed on the actors. We relax this assumption and study a model in which actors play dyadic Prisoner's Dilemmas while also choosing their interaction partners, such that behavior and social networks co-evolve. Can cohesive networks *and* cooperation naturally evolve, is an exogenously imposed cohesive network a condition for cooperation, or are cohesive networks a *result* of high levels of cooperation? We propose formal model in which actors are modeled as boundedly rational and base their decisions to cooperate with a given partner on expectations of the partner's behavior. At the same time, they build or dissolve interactions based on the expected utility of these interactions. Actors form beliefs by learning from their own experience as well as from third-party information obtained via the network. We derive basic analytical results on stable states in this model, and apply computer simulations to study the dynamics of the process in detail. Results suggest that the spread of reputation does not always foster cooperation, and that network cohesiveness is more likely to be a consequence of cooperation than a cause.

1 Introduction

An broadly shared view among social scientists is that social cohesion promotes the emergence of cooperation, trust, and social norms (Homans, 1951; Coleman, 1990; Voss, 2001); a view that is supported by much qualitative (Macaulay, 1963; Greif, 1989, 1994; Ellickson, 1991; Uzzi, 1996, 1997) and some quantitative (e.g., Robinson and Stuart, 2007; Buskens, 2002) evidence. Theoretically, there are various mechanisms that are thought to produce higher cooperation rates in dense networks. Buskens and Raub (2002) make a distinction between *control* and *learning* effects in social networks. Control refers to the idea that actors are more inclined to cooperate if information about defection can spread through the network and lead to sanctions by third parties. Thus, actors would cooperate because they are concerned with the *future consequences* of defection(Raub and Weesie, 1990; Buskens, 2002). A second mechanism is *learning*. Here it is information on *past* interactions that matters. An actor may be more inclined to cooperate with another if she learns from a third party that cooperation with this other actor may be more fruitful than defection, for example, because this other actor appears to be playing according to a Tit-for-Tat strategy. Other useful information that actors may learn through the network concerns the payoffs of the other player (Buskens, 2003; Kreps and Wilson, 1982). While the focus in the literature is mainly on the positive effects of network density on cooperation, the effect may also be negative (Burt and Knez, 1995; Rapoport et al., 1995).

Both control- and learning mechanisms rely on the transfer of information about the behavior of an actor to third parties. For control, this information allows a third party to threaten with sanctions if an actor does not cooperate, while in the learning case this information allows a third party to decide on the expected benefits of cooperation. That is: both mechanisms are examples of *reputation* effects. Reputation, in general, refers to an attribute of an actor ascribed to him by

* Funding for this project was provided by Utrecht University through the High Potentials 2004 subsidy for the research program 'Dynamics of Cooperation, Networks, and Institutions.' Helpful comments and suggestions by Vincent Buskens, Jeroen Weesie, Werner Raub, Noah Mark, and Jacob Dijkstra are gratefully acknowledged.

other actors (Wilson, 1985; Raub and Weesie, 1990). Raub and Weesie (1990) further distinguish between reputation in the narrow sense and reputation in the broad sense. Reputation in the narrow sense refers to situations where the behavior of an actor influences his reputation in *this* situation, while reputation in the broad sense refers to situations where behavior in one interaction influences *other* interactions. We are concerned with reputation in the broad sense. In this paper, we focus only on the learning mechanism of reputation.

Theories on reputation effects commonly assume that social networks are exogenous: the social network is considered as a stable social context which is imposed on the actors, and which provides the means for the spread of information through which the mechanisms of control and learning operate. In recent years, this assumption has been challenged in network research. Increasingly, it is recognized that social networks themselves are also the result of interactions, in which actors make concious decisions about their social relations. This recognition has led to an explosion of interest in social network *dynamics* over the past decades, both in sociology and (even more) in other disciplines such as economics, mathematics and physics (Watts and Strogatz, 1998; Snijders, 2001; Jackson and Watts, 2002a; Dutta and Jackson, 2003; Jackson, 2008).

Relaxing the assumption of static networks might have implications for the predictions about the effects of reputation, for at least two reasons. First, it is not self-evident that reputation mechanisms are equally effective if the network changes as a result of what happens in interactions. For instance, if an actor experiences defection by an interaction partner, she might "spread the word" through her network connections such that the defector can be sanctioned. Alternatively, she might end the interaction with this partner altogether because she prefers not to interact with partners who defect. In that case, she (unintentionally) also changes the possibilities for the spread of reputation for other actors in the network. Conversely, it is possible that network decisions themselves are affected by reputation effects: actors might be more willing to start interactions with potential partners who have a cooperative reputation and less willing to interact with actors who are known to defect. It is theoretically not clear if and how reputation mechanisms function in a dynamic context. The main aim of this paper is to shed some light on this question.

Second, the possibility that networks are dynamic leaves room for a problem of causal order. Traditionally, it is assumed that in the relation between network density and cooperation, the network came first: cooperation is the result of the network structure. As soon as we assume that networks may be dynamic, however, the causal direction could also be in the other direction. If interactions are more likely between actors who tend to cooperate than between actors who defect (for instance, because mutually cooperative interactions are more rewarding than relations of mutual defection), then high density would be the *result* of an already high level of cooperation rather than its cause. Because most empirical studies of network effects on cooperation are cross-sectional, they cannot distinguish between these two possible causal directions.

1.1 Related Theoretical Literature

Models of cooperation in dynamic networks with N-person Prisoner's Dilammas are studied in different setups by Eguíluz et al. (2005), Ule (2005), and Biely et al. (2007). Of these, Ule (2005) includes a reputation mechanism of the control type, but one in which reputation does not depend on the network. Biely et al. (2007) assume some type of reputation by learning, but do not model it explicitly.

Models that study *dyadic* repeated Prisoner's Dilemmas (RPD) (as in our paper) with partner choice can be traced back to (at least) Schuessler (1989), who studies the effects of an "exit-option" in a computational tournament in the style of Axelrod (1984). Vanberg and Congleton (1992), Weesie (1996), Yamagishi et al. (1994), Stanley et al. (1994), and the EdK-Group (2000) all conduct similar analyses in various setups.

None of these studies, however, assume a reputation system: interaction takes place in independent dyads, and although in principle we can speak of a network of interactions, there are no effects of network *structure*. The only study that we are aware of that does discuss a network-based reputation mechanism is by Vega-Redondo (2006), who presents a game-theoretic analysis of the

RPD in an endogenous network. Reputation enters the analysis as players punish interaction partners when they learn of a defection by a partner via the network. Thus, this is reputation in the sense of *control*, which is different from our approach. Moreover, reputation only plays a limited role in network formation. The analysis of this model focusses mainly on the effects of volatility of the environment, and shows that more volatility in the long run leads to lower network density, but also to lower average distance in the network.

Finally, Pujol et al. (2005) study dyadic *support games* in a dynamic network setting with reputation. Although in their model reputation is network-dependent, it's role is limited: third-party information comes only from direct neighbors, and is moreover only used when first-hand experience is not available. A second drawback of their model in our opinion is that actors are modeled as cooperative, which partly "assumes away" the problem we want to address, namely the emergence of cooperation among egoistic actors.

2 The Model

2.1 Formalization of the Problem

The basic interaction is modeled as an infinitely repeated two-person Prisoner's Dilemma. Thus, in every stage of the game, actors play a game as illustrated in Figure 1. In this game, actors can collectively benefit from mutual cooperation, but also have an individual incentive to free-ride on the efforts of the other player. Future payoffs are discounted by a discount parameter w.

	Cooperate	Defect
Cooperate	R, R	S, T
Defect	T, S	P, P

	Cooperate	Defect
Cooperate	$3, 3$	$0, 5$
Defect	$5, 0$	$1, 1$

Fig. 1: A Prisoner's Dilemma Game, in general form and with numerical payoffs, with $T > R > P > S$

There is a finite set of actors $N = \{1, 2, 3, \ldots, n\}$. Actors play two-person Repeated Prisoner's Dilemmas. They can be involved in multiple games at the same time, but can play differently against different partners. For a population of actors, the collection of dyadic relations results in a *network* of relations. A network of n actors can be represented by the $n \times n$ adjacency matrix g, where $g(i, j) = 1$ if there is a link from i to j and $g(i, j) = 0$ otherwise. The network of interactions is undirected by nature, therefore $g(i, j) = g(j, i)$ for all i and j. The set of actors with whom actor i has a link is formally denoted as $N_i(g) = \{j \in N | g_{ij} = 1\}$, and these actors will be referred to as the *neighbors* of i. We uses the terms "link" and "tie" interchangeably.

We make restrictive assumptions of the information available to the actors. Actors are informed of the actions chosen by their interaction partners, but not on the *strategies* for the repeated game according to which their partners are playing. Moreover, actors are not aware of the structure of the *network* beyond their own connections. This also implies that actors are not informed about the payoffs their partners are receiving.

However, while information available to individual actors is rather limited, the network structure allows for diffusion of information among actors, which makes the rise of reputations possible. But also the diffusion of information is assumed to have its limits: it does not flow effortlessly through the network but decreases in reliability the further it travels through the network. We model this process explicitly when we discuss how actors use information.

A key property of our model is that the interaction network is not exogenously imposed, but can be changed by the actors. Because network ties represent interactions, we assume that ties can only be created with mutual consent, while they can be deleted unilaterally (cf. Jackson and Wolinsky, 1996).

An important assumption is that maintaining network ties is costly. The underlying reasoning for this assumption is that maintaining social interactions requires some effort, and that therefore actors may want to end relations that are less profitable. Assigning a fixed cost to every interaction is a convenient way to model this (cf. Jackson and Watts, 2002b).

Formally, the total cost k for maintaining z ties for actor i in each round of play is given by the simple linear function

$$k_i(g) = \alpha z_i \tag{1}$$

in which z_i denotes the number of ties i is involved in (i.e., $z_i = |N_i|$) and $\alpha \geq 0$.

The various components of the model are combined into a *repeated* game in the following way. Each period of the process consists of three phases: In the *Network formation phase*, a number of pairs of actors (denoted by η) are randomly chosen to update their relation (create a new tie if their is none, or remove a tie if there is one), using the information available to them. In the *Game play phase*, all actors simultaneously play the PD with each of their neighbors, given the network resulting from phase 1. Lastly, in the *Information phase*, actors are informed of the outcomes of the game play phase, information spreads through the network (by the mechanism we explain below), and actors update their beliefs about each other.

After the information phase, we return to the network formation phase and the process continues. In phases 1 and 2, actors rely only on the information they received in phase 3, from own experience or through reputation. Thus, after new ties are formed in phase 1, there is no additional spread of information, and the decision in the game is based on the same information as the decisions in the network formation phase.

2.2 Individual Strategies

In summary, the decision situation for the actors can be characterized as a complex repeated game with incomplete information, in which the information structure is changed endogenously. This leads to an extremely complex decision situation with a very large range of action alternatives and possible outcomes. This situation would put such extreme requirements on a truly rational actor in terms of information processing and computation that we think it unlikely that human actors would be capable of acting rationally in such a context.

Instead, we prefer to model actors as *boundedly rational*, in the sense that actors make a number of simplifying assumptions about the world around them and make use of information in a possibly suboptimal manner (Simon, 1956; Rubinstein, 1998). More specifically, we assume that actors consider only "$t-1$ matching strategies" by the opponent (Downing, 1975). That is, actors assume that their opponent's action in the game is a response to their own action in the game in the previous period.

$p(C_j \mid C_i)_t$**:** The probability that j will cooperate at time t after i cooperated at time $t-1$;
$p(C_j \mid D_i)_t$**:** The probability that j will cooperate at time t after i defected at time $t-1$;

We propose that actors try to maximize utility in the repeated game by assuming to play against an opponent who uses a $t-1$ matching strategy with unknown response probabilities.[3] As the precise strategy of the opponent is not revealed, maximizing utility will involve making an assessment of the most likely values of $p(C_j \mid C_i)$ and $p(C_j \mid D_i)$ for the strategy of the opponent. In effect, this means that actors assume to be playing against a probabilistic *automaton*, which is driven by two conditional probabilities. This leaves the actor with two tasks: first, to determine the probabilities according to which the opponent is playing, and subsequently, to determine the optimal response given these probabilities.

To assess the behavior of the opponent in terms of $p(C_j \mid C_i)$ and $p(C_j \mid D_i)$, actors simply use the frequency distribution of the opponent's responses in the game so far. That is:

$$p(C_j \mid a_i)_t = \frac{C_{ijt}(a)}{T_{ijt}(a)} \tag{2}$$

[3] This approach is closely related to *fictitious play* (Fudenberg and Levine, 1998).

in which $T_{ijt}(a)$ is the total number of times that i played action a (C or D) against j until time t, and $C_{ijt}(a)a$ is the total number of times that j reacted with cooperation on action a by i until time t.

The next task is to determine the optimal response when playing against an automaton with two given response probabilities $p(C_j \mid C_i)$ and $p(C_j \mid D_i)$. For $w \to \infty$, which we'll assume throughout, it can be shown that to maximize utility, it is sufficient to consider only three different courses of action, which we'll call *substrategies* for convenience. An actor can cooperate in every round (labeled as ALLC), defect in every round (labeled as ALLD), or alternate between cooperation and defection (labeled as ALT).

The actor's behavior as described so far is equivalent to the strategy for the Repeated Prisoner's Dilemma known as "DOWNING", which was one of the competitors in the famous computer tournament conducted by Axelrod (1984), and which was originally proposed by Downing (1975). From here on we will use the label "DOWNING" to refer to the individual strategy used in this paper.

2.3 Reputation

The reputation of actor j with actor i (that is, the information that i has about the behavior of j in interactions not with i) consists of information of j's responses against her neighbors. We assume that information received by third parties has a smaller influence on i's decision than her own experience. For instance, i might have less confidence in third party information because it is more likely to be distorted the farther it travels through the network. To model such effects, we assume that the weight of the third-party information that i receives about j depends on the *network distance* δ through which this information is transfered. If i and j are connected, then the reputation of j with i consists of the expectations of j's neighbors k about j weighted by the network distance δ_{ik} between i and every k. δ_{ik} is defined as the *shortest path* through the network between i and k (see Wasserman and Faust, 1994, 110). The information is weighted in the sense that the information obtained from k is subject to network *decay*, and is considered less important to i if the shortest distance from k to j is larger. These quantities, in combination with those obtained from own experience, are used to compute the probabilities that DOWNING uses to assess the opponent's future behavior. To include reputation we modify equation (2) in the following way. Recall that N_i denotes the set of j's neighbors. Then, the probability $p(C_j \mid a_i)$ that j will cooperate at time t after i played action a is defined as

$$p(C_j \mid a_i)_t = \frac{C_{ijt}(a) + \sum_{k \in N_j(g_t)} \omega^{\delta_{ikt}} C_{kjt}(a)}{T_{ijt}(a) + \sum_{k \in N_j(g_t)} \omega^{\delta_{ikt}} T_{kjt}(a)} \tag{3}$$

in which ω ($0 \leq \omega \leq 1$) is the extent of network decay of information. Thus, if $\omega = \frac{1}{2}$, information learned via reputation is discounted with a factor $\frac{1}{2}$ with every step it travels through the network. With $\omega = 0$, reputation does not play any role; with $\omega = 1$ information from any source is given the same weight.

2.4 Network Decisions

When making linking decisions, actors only maintain (or create) a tie if the expected payoff from the interaction exceeds the cost of maintaining the tie. The expected payoff is the payoff the actor would obtain *if* he would interact with the potential partner under consideration. If the result of this evaluation is nonnegative for *both* actors, the tie is created (or kept); if the result is negative for one (or both) of the actors it is not created, or removed.

Because we want to keep the focus of the paper on reputation effects in dynamic networks, we abstract here from strategic considerations concerning the transmission of information: we just assume that information flows through the network, without modelling decisions of actors to actually pass this information. Lippert and Spagnolo (2005) discuss this issue , but abstract from

network formation. Similarly, we assume that actors do not purposively try to obtain strategic information.[4]

We define the dynamic process as converged if two conditions are met. The first condition is that there is no pair of actors willing to create a new tie, and no single actor willing to remove one tie. This criterion conforms to the notion of *pairwise stability* (Jackson and Wolinsky, 1996). The second condition is that the beliefs of actors are stable, in the sense that they converge.The convergence of beliefs implies that substrategies are also stable.

3 Analysis of the Model

The main focus of the formal analysis is on *stable states* of the co-evolution process, that is, converged states of the process. First, we briefly discuss the behavior of the DOWNING strategy in the two-person case, when two players using the DOWNING strategy are interacting, without the possibility to end the interaction. This will be helpful in the subsequent analysis for static and dynamic networks. The main result on the two-player case can be summarized as follows:

Theorem 1. *If two actors are both using the* DOWNING *strategy in a two-person RPD, the only stable substrategy-combinations are (*ALLC,ALLC*) and (*ALLD,ALLD*).*

Due to space restrictions, we omit the proof here, given that this result is only instrumental to the main results of the paper. Central to the argument is showing that no strategy combination that includes ALT can be stable.

3.1 Stable States in Static Networks

We now extend the analysis to the "networked" setting in which actors play with multiple partners, but first only consider *static* networks ($\eta = 0$). The crucial difference between the network setting and the two-player setting analyzed in the previous section is that actors can share information, and reputations can arise. In the model, the extent to which reputation plays a role is determined by the parameter ω. We shall first look at two extreme cases, namely $\omega = 0$ (no diffusion of reputations) and $\omega = 1$ (perfect diffusion of reputations). Let σ_{ij} denote the substrategy used by i against a neighbor j; σ_{ij} can be either ALLC, ALLD, or ALT.

Theorem 2.

(i) If $\omega = 0$ and $\eta = 0$, then $\forall j \in N, \forall i \in N_j(g), \sigma_{ij} = \sigma_{ji} \wedge (\sigma_{ij} = \text{ALLC} \vee \sigma_{ij} = \text{ALLD})$ in stable networks.

(ii) If $\omega = 1$ and $\eta = 0$, then $\forall j \in g, \forall i, k \in N_j(g), \sigma_{ij} = \sigma_{kj}$ in stable networks.

Proof. Case (i) of Theorem 2 is simply a reiteration of Theorem 1. In the absence of reputation effects ($\omega = 0$) the network setting is equivalent to the two-person setting. If $\omega = 1$ (case (ii)), then equation (3) reduces to

$$p(C_j \mid a_i) = \frac{C(a)_{ij} + \sum\limits_{k \in N_j(g)} C(a)_{kj}}{T(a)_{ij} + \sum\limits_{k \in N_j(g)} T(a)_{kj}} \tag{4}$$

(omitting the subscript t) such that $p(C_j \mid a_i) = p(C_j \mid a_k)$ for all i, j, k who are directly or indirectly connected. This implies if two actors are (directly or indirectly) connected, they are acting on the same information. Because the choice of a substrategy depends exclusively on the conditional probabilities of cooperation, it follows that all i and k will choose the same substrategy against j. □

[4] We hope to relax this assumption in future developments of this model.

Theorem 2 states that in the absence of reputation effects, interaction in static networks does not differ from interaction in isolated dyads, and behavior converges to mutual ALLC or mutual ALLD. If reputation effects are at work, however, behavior within dyads is no longer necessarily symmetric because the observations on the behavior of a partner can be "compensated" by observations of his behavior in interactions with other actors. In the extreme case of perfect information transmission (case (ii)), all partners of a given actor will choose the same strategy against this actor, but this does not imply that all actors choose the same behavior *against all their partners.* Examples can be constructed in which a group of actors cooperate with each other, but defect against a single actor who in turn cooperates with all of them. In this case, the defections of these actors are "offset" by the observation that they cooperate in all their other relations.

3.2 Stable States in Dynamic Networks

Lastly, we turn to the situation in which games are played in *dynamic* networks. The important difference as compared to the static network case is that certain types of interactions will be discontinued if the expected reward from the interaction is less than the cost of maintaining the relation. We distinguish between two different scenarios with regard to the cost of network formation: the case where $T > R > \alpha > P > S$ and the case where $T > R > P > \alpha > S$, where α is the "maintenance cost" of a tie. In the first case, only interactions in which cooperation takes place are attractive. In the second case, also relations with mutual defection are attractive, but actors will prefer isolation over exploitation ($\alpha > S$). Let $g*$ denote a *component* of network g, that is: a subnetwork of g, consisting of a maximal subset of nodes and links such that all nodes are directly or indirectly connected.

Theorem 3.

(i) If $\omega = 0$ and $R > \alpha > P$, then $\forall i \in N, \forall j \in N_i(g), \sigma_{ij} = \text{ALLC}$ in stable networks. Any network configuration is possible.

(ii) If $\omega = 0$ and $P > \alpha > S$, then the network is complete and $\forall i, j \in N, \sigma_{ij} = \sigma_{ji} \wedge (\sigma_{ij} = \text{ALLC} \vee \sigma_{ij} = \text{ALLD})$ in stable networks.

(iii) If $\omega = 1$ and $R > \alpha > P$, then the network may consist of one or more complete components $g*$, while actors may also be isolated. $\forall g* \subset g, (\forall i \in N(g*), \forall j, k \in N_i(g*), \sigma_{ji} = \sigma_{ki}) \wedge \neg(\forall i \in N(g*), \forall j \in N_i(g*), \sigma_{ij} = \text{ALLD})$.

(iv) If $\omega = 1$ and $P > \alpha > S$ all links are present in the network and $\forall i \in N, \forall j, k \in N_i(g) \sigma_{ji} = \sigma_{ki}$.

Proof. Case (i) is a reiteration of case (i) of Theorem 2, with the addition that interactions in which both actors use ALLD are no longer stable as $\alpha > P$ (note that the expected payoff per round in that case converges to P). Case (ii) differs from case (i) in that interactions in which both actors use ALLD are also stable, as the expected payoff in these interactions converges to P and $P > \alpha$. Moreover, since P is the minimal expected payoff, all links are created. Case (iii) partly relies on the same argumentation as case (ii) of Theorem 2: all actors who are directly or indirectly connected will base their choices against a given actor i on the same information. Therefore, *within* a component $g*$, all links must be present, because if it is profitable for any actor i to connect to j, it must be profitable to connect to j for all actors who are directly or indirectly connected to i. For the same reason, all neighbors of j use the same substrategy against j. It is not possible that $\sigma_{ij} = \sigma_{ji} = \text{ALLD}$ for all $i, j \in g*$. In that case, $p(C_j|D_i)$ would go to zero for all $i, j \in g*$ and the expected payoff of all interactions would go to P, which is lower than α. However, it is possible to construct examples such that some of the actors in a component play ALLD and others is play ALLC. Case (iv) differs from case (iii) in that because $\alpha < P$ and P is the lowest possible payoff, all links must be created such that any stable network must be complete. This implies that the situation in which $\sigma_{ij} = \sigma_{ji} = \text{ALLD}$ for all $i, j \in g$ is also stable. □

Theorem 3 states that if the cost of tie maintenance is low enough (cases (ii) and (iv)), the complete network will form. If the cost of tie maintenance is high, either any network configuration is possible (case (i)), or the network will consist of fully connected components (case (iii)).

Perfect transfer of information (cases (iv) and (iii)) has the effect that components must be fully connected. Next, let us compare the situation without reputation ($\omega = 0$) with the situation with full reputation ($\omega = 1$). When costs are low (cases (ii) and (iv)), the complete network will form and there are no clear effects of reputation. When costs are high however (cases (i) and (iii)), we see that the presence of reputation effects clearly has an effect on the possible stable distribution of behavior. Without reputation (case (i)), only mutually cooperative ties are stable. *With* reputation (case(iii)), constellations in which some actors defect are also stable, as long as it is not the case that *all* actors in a component play ALLD against all their neighbors. Thus, we see that, with high cost of relations, reputation opens the door for defection. However, this does not necessarily mean that reputation always leads to lower overall cooperation. While case (i) states that only cooperative interactions are stable, the emerging network of interactions may be very sparse or even empty, yielding a very low level of cooperation. The question whether processes with and without reputation effects lead to different levels of overall cooperation will be addressed in the computer simulations in the next section.

The theorems characterize stable networks for the extreme cases in which $\omega = 0$ or $\omega = 1$, but not for intermediate values of ω. In many cases the characterizations are rather general, and allow for many constellations of network and behavior to be stable. Moreover, the theorems do not provide any insight into which of the possible stable states are more or less likely, given some present state of the process. That is, the analyses say nothing about the *dynamic process* which brings about stable states. In the following sections, we conduct computer simulations to address these issues.

4 Setup of the Simulation

We run computer simulations of the model to study effects of intermediate values of ω, and to study the effects of initial conditions in the dynamic process. We vary ω, η, α, the initial network structure, and the initial tendency for cooperation (λ). For the payoffs of the game we choose $T = 5, R = 3, P = 1$ and $S = 0$. We vary the extent to which reputation travels through the network, ω, between 0 and 1. We vary the speed of network formation η to be zero ($\eta = 0$; the network is static), relatively slow ($\eta = 30$, the number of actors), or relatively fast ($\eta = 435$, the number of dyads). $\eta = 0$ refers to the situation in which networks are static. Linear cost of ties α are chosen such that $S < \alpha < P$, $P < \alpha < \frac{T+S}{2}$, or $\frac{T+S}{2} < \alpha < R$. These values are chosen such that, in the first case, the cost of tie maintenance allows for relations with mutual defection, mutual alternation and mutual cooperation, in the second case for only mutual alternation and mutual cooperation, and in the third case only for mutual cooperation.

The initial conditions of the simulation consist of the initial beliefs of the actors and the initial network structure. Parameter λ governs the initial distribution of beliefs, in such a way that the higher λ, the larger $p(C_j \mid C_i)_{t_0}$ relative to $p(C_j \mid D_i)_{t_0}$, and the higher the overall tendency for cooperation. As a larger λ leads to an —on average— higher expectation that opponents will be reactive, λ might also informally be interpreted as "optimism".

For the initial network structure, we draw from a set of artificially generated network structures. To construct these networks, we use various well-known network models, including the Erdős-Renyi random graph model, the small-world model of Watts and Strogatz (1998) and the preferential attachment model of Barabási and Albert (1999). These models have been shown to reproduce some key characteristics of empirical networks, and the resulting networks therefore provide a reasonably plausible set of initial networks for the simulation.[5] We vary the parameters of the respective generating algorithms in such a way to obtain a reasonable variance in network density and network centralization. Table 1 summarizes the initial conditions used in the simulation.

Two types of outcomes are of interest for our inquiry: those related to cooperation and those related to the emerging network. To express the amount of cooperation in the process, we define

[5] The results do not differ between different network generating algorithms, which we take as an indication that the precise method used does not matter much, and that studying additional methods is not likely to yield new insights.

two different measures. The first is *the proportion of cooperation in interactions*, that is, among connected actors. This measure is conditional on the emerging network: if the network is very sparse, the level of cooperation *within* this network can still be high even though very few actors are actually cooperating. For an empty network, this measure is not defined. Second, we study *the total proportion of cooperation*, that is, the proportion of cooperative choices of all possible interactions. The measure is 1 if all actors are connected *and* cooperate in all interactions, and is 0 if there is no cooperation in any interaction, or if there are no interactions at all (the network is empty). This measure is less comparable across different network structures because its maximal value is restricted by network density, but gives a better indication of overall welfare. We use the proportion of cooperation per interaction (or cooperation per *tie*) to study emerging cooperation in static networks. For dynamic networks, total cooperation is a more suitable measure than cooperation per interaction because the number of interactions is itself an outcome of the process. Although these two measures may look very different at first sight, note that when one is applied to static networks and the other to dynamic networks, they both measure the *proportion of maximally attainable cooperation*. Thus, we believe it is justified to use these two measures to compare cooperation between static and dynamic networks.

As a convergence criterion for the simulation, we require that the largest change in the beliefs of all actors is smaller than 5%.

Table 1: Initial conditions in the simulation

	Mean	SD	Min	Max
Density	0.50	0.31	0.00	1.00
Centralization	0.06	0.08	0.00	0.33
λ	2.50	1.12	1.00	4.00

5 Simulation results

The results reported here come from on a total of 7200 simulation runs. In the runs with a static network ($\eta = 0$), the process always converged within 1000 rounds. Of the runs on a dynamic network ($\eta > 0$), 99.8% within 1000 rounds. In the results that follow, we include only runs that converged within 1000 rounds.

5.1 Results on static networks

We here only sketch the results on static networks.[6] We find that cooperation depends heavily on λ. This result indicates that the initial conditions of the process have a strong impact on the outcomes. We also see effects of reputation diffusion on the outcomes. While the outcomes lie close together when reputation effects are absent ($\omega = 0$), the variance in cooperation tends to increase with ω. This implies that the presence of reputation effects allows for more extreme levels of cooperation, both high and low. The exception to this trend is the case where we $\lambda = 4$, which is the most favorable condition for cooperation. Here, strong reputation effects *decrease* the variance in cooperation outcomes, especially in dense networks. If the density is smaller than .75, a small reputation effect ($\omega = .25$) allows for lower levels of cooperation.

5.2 Results on dynamic networks

Table 2 shows average cooperation levels and network density for different combinations of values of α (cost of ties) and ω (reputation). For $\alpha = 0.9$, all stable networks are complete because the

[6] More detailed results are available from the first author.

cost of maintaining a tie is lower than P, the minimally expected payoff (cf. Theorem 3, cases (ii) and (iv)). Net effects of reputation on cooperation are marginal with this cost level.

If $\alpha = 1.9$, the cost of maintaining a tie is higher than P, which makes mutually defective relations unstable in the absence of reputation effects. Indeed, cooperation per interaction is almost 100% if $\omega = 0$.[7] Cooperation per dyad, however, is comparable to the lower cost regime. When $\omega > 0$, the density of stable networks increases, while the level of cooperation per interaction remains more or less constant. That is: there are more interactions, but these are not cooperative interactions. Thus, in line with the analytical results (Theorem 3), when reputations are allowed to spread through the network, this allows for defection to survive even if the cost of maintaining a relation are higher than the expected payoff from a mutually defective relation. As a consequence, the level of cooperation over all dyads decreases if $\omega > 0$. These patterns are, however, not linear in ω: density jumps when ω increases from 0 to .25, but then decreases. Similarly, the drop in cooperation per dyad is largest between $\omega = 0$ and $\omega = .25$, and much smaller between higher levels of ω. When the cost of a tie is only slightly lower than the expected payoff from a mutually cooperative relation ($\alpha = 2.9$), we see that without reputation effects ($\omega = 0$), cooperation in interaction is maximal, but density is much lower than with lower cost. When $\omega > 0$, we again see that cooperation per interaction decreases. At the same time, density decreases with ω, until almost 0 when $\omega = 1$. In this case, cooperation per interaction is still high, but there are very few interactions, such that the network is extremely sparse.

Given the strong effects of the initial tendency for cooperation that we found for static networks, we should also compare the outcomes for different values of λ. Figure 2 shows average cooperation *per dyad* as depending on the strength of the reputation mechanism (ω), for different values of α and λ. As compared to the case of static networks, we generally find a somewhat stronger effect of ω. The direction of this effect depends heavily on α and λ. For the two lower values of α the, the effect of ω is negative for lower values of λ and positive for higher values of λ. Thus, on average, the spread of reputation "catalyzes" the tendency the process already had at its start. This does not mean, however, that a strong reputation effect leads to high cooperation with a high λ and low cooperation with low λ. As in static networks, the range of stable outcomes also increases with ω. For $\alpha = 2.9$, the result is different in a number of respects. First, we see that the effect of ω on cooperation is nearly always negative (or zero). This is even the case for the highest value of λ, which with lower costs mostly leads to full cooperation. Second, we see an interesting *divergence* of outcomes when $\lambda = 4$ and $\omega = .25$. Here, a number of simulation runs converged on an average lower level of cooperation as compared to the situation where $\omega = 0$, while another group of runs converged on significantly *higher* levels of cooperation. Closer analyses of these latter cases reveal that they are characterized by a relatively lower network density as an initial condition. Among all the runs with $\lambda = 4$, $\alpha = 2.9$, and $\omega = .25$, the correlation between initial density and cooperation is .59. An explanation of these results might be that if the network is initially sparse, limited diffusion of reputation helps to form a network of cooperative relations. If the network is dense from the start, in contrast, the diffusion of reputation mostly serves to "spread bad news", which prevents the further buildup of a cooperative network.

Lastly, we study the effects of the *initial* density of the network in dynamic networks. Figure 3 presents scatterplots of average cooperation per dyad per ω, λ and α as in Fig. 2, separately for different initial network densities. To reduce the number of graphs we take density rounded on multiples of 0.25, and show only values of λ for which we find interesting differences. The figure shows an interesting interaction effect between initial density and ω. For lower values of α, the effect of ω is clearly stronger for lower initial density. Moreover, we see that the variance of stable states increases more strongly with ω, especially for $\lambda = 3$. This means that when the initial density is higher, higher levels of cooperation can be reached with lower levels of ω. For the highest value of $\alpha <$, the effects are less clear. For $\lambda = 3$, the effect of omega becomes stronger for *higher density*, while for $\lambda = 4$, there is no clear interaction effect. We can however identify the "special cases" mentioned in the previous section, where an exceptionally high level of cooperation is reached for

[7] The proportion of cooperation is not perfectly 1, because in the simulation, convergence may be imperfect .

Table 2: Average density and cooperation in stable networks by α and ω

α		ω 0	.25	.5	.75	1
0.9	Coop. per interaction	0.45	0.48	0.47	0.46	0.45
	Coop. per dyad	0.45	0.48	0.47	0.46	0.45
	Density	1.00	1.00	1.00	1.00	1.00
1.9	Coop. per interaction	0.92	0.64	0.59	0.59	0.53
	Coop. per dyad	0.43	0.46	0.42	0.43	0.43
	Density	0.44	0.68	0.63	0.61	0.62
2.9	Coop. per interaction	1.00	0.99	0.97	0.95	0.89
	Coop. per dyad	0.18	0.11	0.10	0.04	0.02
	Density	0.18	0.11	0.10	0.04	0.02

$\omega = \frac{1}{4}$. The implication of these results is that the combination of an initially high density and the presence of reputation effects is *not* the best recipe for cooperation if the network is dynamic. On the contrary: if there is a positive effect of reputation, this effect is most pronounced when the initial density is low.

6 Conclusions and Discussion

The overall results can be summarized as follows. First, we find that if networks are exogenously determined, the range of possible stable states increases with the extent of reputation diffusion and the density of the network. States with higher overall cooperation levels emerge as compared to situations with less reputation diffusion, but also states with *lower* cooperation rates. Thus, we do *not* find that reputation effects through networks always lead to more cooperation, as is most commonly assumed in the sociological literature. Rather, we find that relatively higher cooperation is a possible consequence of reputation effects, but so is lower cooperation. These findings are in line with by Burt and Knez (1995) who argue that reputation effects generally lead to more extreme outcomes. However, while Burt and Knez (1995) rely on psychological mechanisms to explain this phenomenon, we show that it can also emerge from a simple learning model.

Second, we find that if the network is dynamic, the spread of reputation on average tends to "catalyze" the initial tendency of the process towards higher or lower levels of cooperation. Moreover, we find strong interaction effects of the cost of maintaining ties and reputation effects in dynamic networks. When the cost of a tie becomes very high, maintaining a network of cooperative relations becomes difficult, and the addition of reputation makes this still worse, leading to empty networks in many cases.

Third, we find no indications that, in a context in which the network structure is endogenous, high cooperation levels are likely to be the result of reputation effects in an initially dense network. Instead, we find that, in dynamic networks, the effect of the spread of reputations tends to be stronger if the network is *less* dense. That is: the diffusion of reputation is most likely to lead to high cooperation rates if initial beliefs are "optimistic" (λ is high) and the network is sparse. As a *result* of high cooperation, a dense network emerges. An interpretation of this effect could be that if the network is initially sparse, actors have the opportunity to initiate interactions only with those partners whom they expect to act cooperatively. The diffusion of reputation then helps in the further buildup of this "cooperative network." If, in contrast, the network is dense from the beginning, there will also be some relations in which actors are dot not behave cooperatively. In this case, the diffusion of reputation only helps to spread to "bad news", which hinders the development of cooperation.

While we believe that our analysis adds new insights to the study of cooperation in networks, the model also has some limitations. First, we modeled reputation only as learning, and did

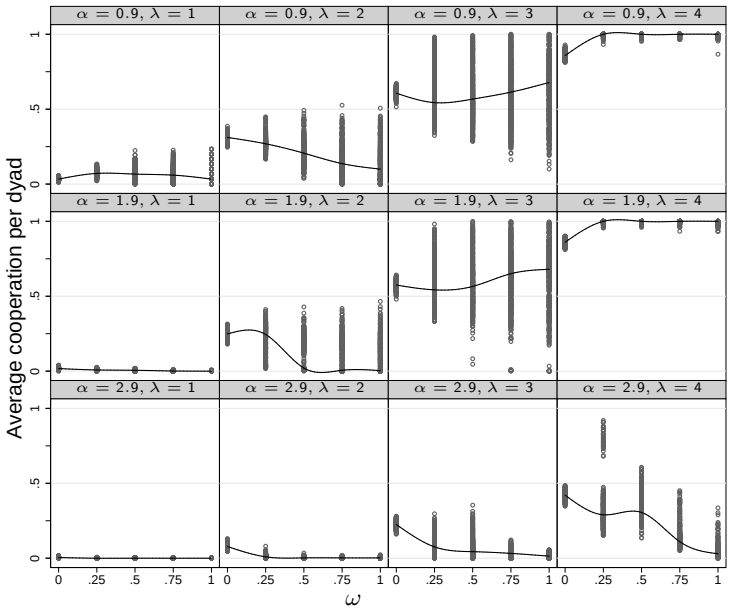


Fig. 2: Average cooperation per dyad by ω in dynamic networks, with median splines added: graphs by λ (rows) and α (columns)

not take into account that actors may care about their *future* reputation. Adding such "control" mechanisms to the model however would not only make the model much more complex to analyze, but would also put considerably higher demands on actors' rationality.

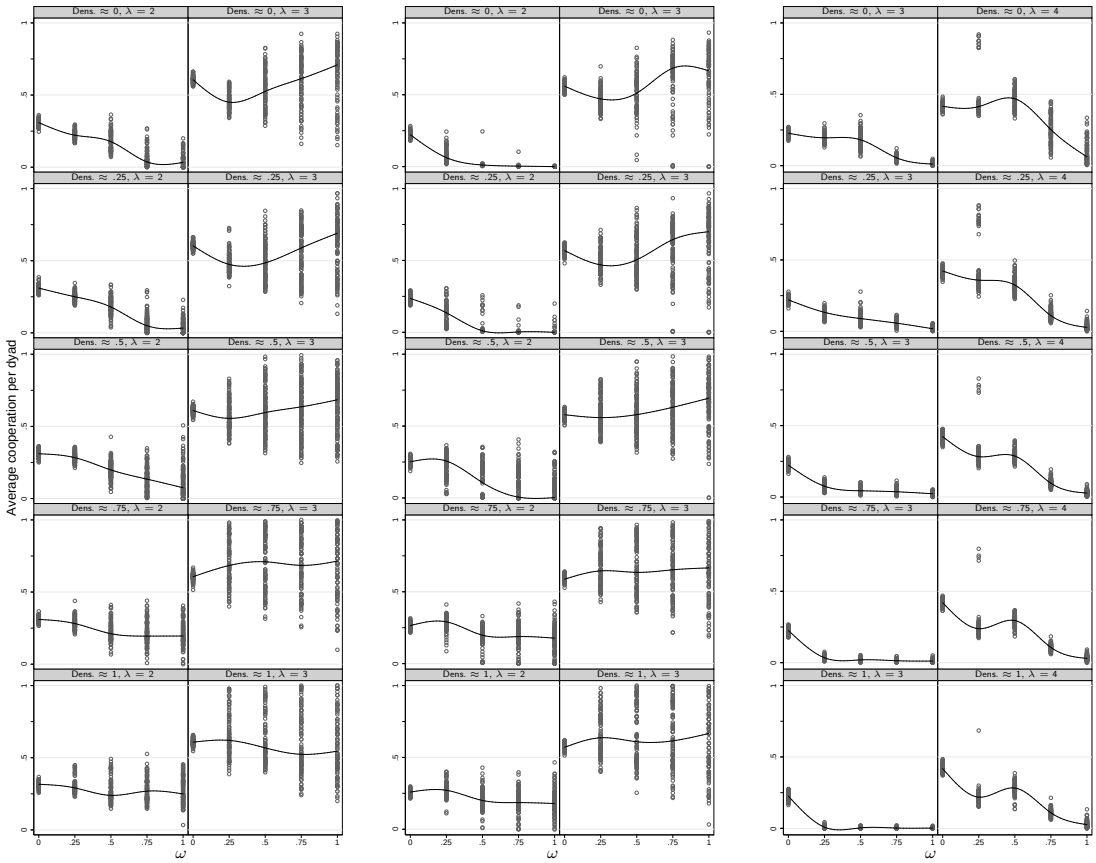
Second, we did not model actors' expectations on the *linking* behavior of their interaction partners. In effect, actors in our model assume that their opponents will never unilaterally end the interaction. We did so because we wanted to focus strictly on reputation effects for this paper, but the model could be extended in this direction by including expectations about the opponent's linking behavior as conditional probabilities assigned to the beliefs of the actor.

Third, we did not consider information diffusion as a strategic choice. For simplicity, we assumed that the transfer of information is automatic, and actors do not make an explicit decision about whether or not to pass along specific information. For many empirical applications, this assumption is unrealistic: as we can learn from studying gossip (Gambetta, 1994; Burt, 2001) people often have reasons to think strategically about what to tell to whom.

Fourth, the learning model applied here is rather simple: actors have very simplified expectations about their opponent's behavior, and update those beliefs using very basic methods. More complex learning models are conceivable, in which actors, for example, assume that their partners' behavior is conditional on a longer history than only the previous round, or take the reliability of their estimations into account.

Finally, it would be interesting to look at *stochastic* stability of the process. At present, our model is basically deterministic; the only stochastic element is the order in which actors update their behavior. A stochastic approach might help to reduce variability of predicted stable states (cf. Jackson and Watts, 2002b), and substantively, give insights in how reputation affects cooperation and network formation in a volatile environment. A first intuition is that the introduction of noise would make cooperation even more fragile, but that the spread of reputations could help to "protect" group cohesion and cooperation from small mistakes, as they are compensated by the good reputation of an actor.

Broadly speaking, we see two ways to develop the model further. The first is to address some of the theoretical issues mentioned above as extensions of the model. Another question is, of course, the *empirical* validity of the model. Eventually, we aim to explain and predict real empirical



(a) $\alpha = 0.9$ **(b)** $\alpha = 1.9$ **(c)** $\alpha = 2.9$

Fig. 3: Effects of reputation on cooperation by initial density, λ and α

processes of cooperation and network formation with our model. At present, however, with so many theoretical issues still unresolved, testing predictions from the model directly with real-world data is likely to run into problems. For instance, if discrepancies between predictions and the data are found, it will be difficult to determine whether the discrepancy is caused by the overly simplified assumptions about actors' decision making, or by the misspecification of the underlying game. Probably, a more fruitful approach to test the model empirically would be to conduct controlled experiments. Using the methods developed in experimental economics and social psychology, one can study human behavior in complex strategic interactions "at close range", while controlling properties of the larger environment. Such an approach would be most useful to assess the extent to which the model's assumptions about actors' decision making are sufficient or are in need of modification.

Bibliography

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Barabási, Albert-László and Réka Albert. 1999. "The Emegence of Scaling in Random Networks." *Science* 286:509–512.

Biely, Christoly, Klaus Dragostis, and Stefan Thurner. 2007. "The Prisoner's Dilemma on Co-Evolving Networks under Perfect Rationality." *Physica D* 228:40–48.

Burt, Ronald S. 2001. "Bandwidth and Echo: Trust, Information, and Gossip in Social Networks." In *Networks and Markets: Contributions from Economics and Sociology*, edited by Alessandra Casella and James E. Rauch, pp. 30–74. New York: Russell Sage Foundation.

Burt, Ronald S. and Marc Knez. 1995. "Kinds of Third-Party Effects on Trust." *Rationality and Society* 7:255–292.

Buskens, Vincent. 2002. *Social Networks and Trust*. Boston, MA: Kluwer Academic Publishers.

Buskens, Vincent. 2003. "Trust in Triads: Effects of Exit, Control, and Learning." *Games and Economic Behavior* 42:235–252.

Buskens, Vincent and Werner Raub. 2002. "Embedded Trust: Control and Learning." In *Group Cohesion, Trust and Solidarity*, edited by Shane R. Thye and Edward J. Lawler, volume 19 of *Advances in Group Processes*, pp. 167–202. Amsterdam: Elsevier Science.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge, MA: Belknap.

Downing, Leslie. 1975. "The Prisoner's Dilemma Game as a Problem-Solving Phenomenon." *Simulation & Games* 6:366–391.

Dutta, Bashkar and Matthew O. Jackson (eds.). 2003. *Networks and Groups. Models of Strategic Formation*. Heidelberg: Springer-Verlag.

EdK-Group. 2000. "Exit, Anonimity and the Chances of Egoistical Cooperation." *Analyse und Kritik* 22:114–129.

Eguíluz, Victor M., Martin G. Zimmerman, and Camilo J. Cela-Conde. 2005. "Cooperation and the Emergence of Role Differentiation in the Dynamics of Social Networks." *American Journal of Sociology* 110:977–1008.

Ellickson, Robert C. 1991. *Order without Law. How Neighbors Settle Disputes*. Cambridge, MA: Harvard University Press.

Fudenberg, Drew and David K. Levine. 1998. *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Gambetta, Diego. 1994. "Godfather's Gossip." *Archives Européennes de Sociologie* 35:199–223.

Greif, Avner. 1989. "Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders." *Journal of Economic History* 49:857–882.

Greif, Avner. 1994. "Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies." *The Journal of Political Economy* 102:912–950.

Homans, George C. 1951. *The Human Group*. London: Routledge.

Jackson, Matthew O. 2008. *Social and Economic Networks*. Princeton, NJ: Princeton University Press.

Jackson, Matthew O. and Alison Watts. 2002a. "The Evolution of Social and Economic Networks." *Journal of Economic Theory* 106:265–295.

Jackson, Matthew O. and Alison Watts. 2002b. "On the Formation of Interaction Networks in Social Coordination Games." *Games and Economic Behavior* 41:265–291.

Jackson, Matthew O. and Asher Wolinsky. 1996. "A Strategic Model of Social and Economic Networks." *Journal of Economic Theory* 71:44–74.

Kreps, David M. and R. Wilson. 1982. "Reputation and Imperfect Information." *Journal of Economic Theory* 27:253–279.

Lippert, Steffen and Giancarlo Spagnolo. 2005. "Networks of Relations and Social Capital." SSE/EFI Working Paper Series in Economics and Finance, Stockholm School of Economics.

Macaulay, Steward. 1963. "Non-Contractual Relations in Business: A Preliminary Study." *American Sociological Review* 28:55–67.

Pujol, Josep M., Andreas Flache, Jordi Delgado, and Ramon Sangüesa. 2005. "How Can Social Networks Ever Become Complex? Modelling the Emergence of Complex Networks from Local Social Exchanges." *Journal of Artificial Societies and Social Simulation* 8.

Rapoport, Anatol, Andreas Diekmann, and Axel Franzen. 1995. "Experiments with Social Traps IV: Reputation Effects in the Evolution of Cooperation." *Rationality and Society* 7:431–441.

Raub, Werner and Jeroen Weesie. 1990. "Reputation and Efficiency in Social Interactions: An Example of Network Effects." *American Journal of Sociology* 96:626–654.

Robinson, David T. and Toby E. Stuart. 2007. "Network Effects in the Governance of Strategic Alliances." *Journal of Law, Economics and Organization* 23:242–273.

Rubinstein, Ariel. 1998. *Modeling Bounded Rationality*. Cambridge, MA: MIT Press.

Schuessler, Rudolf. 1989. "Exit Threats and Cooperation under Anonymity." *Journal of Conflict Resolution* 33:728–749.

Simon, Herbert A. 1956. "Rational Choice and the Structure of the Environment." *Psychological Review* 63:129–138.

Snijders, Tom A. B. 2001. "The Statistical Evaluation of Social Network Dynamics." In *Sociological Methodology*, edited by Michael E. Sobel and Mark P. Becker, volume 31, pp. 361–395. Boston, MA: Blackwell.

Stanley, E.A., D. Ashlock, and M.D. Smucker. 1994. "Iterated Prisoner's Dilemma with Choice and Refusal of Partners: Evolutionary Results." In *Artificial Life III*, edited by Christopher G. Langton, volume XVII of *Santa Fe Institute Studies in the Sciences of Complexity*. Addison-Wesley.

Ule, Aljaž. 2005. *Exclusion and Cooperation in Networks*. Amsterdam: Thela Thesis.

Uzzi, Brian. 1996. "The Sources and Consequences of Embeddedness for the Economic Performance of Organizations: The Network Effect." *American Sociological Review* 61:674–698.

Uzzi, Brian. 1997. "Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness." *Administrative Science Quarterly* 42:35–67.

Vanberg, Viktor J. and Roger D. Congleton. 1992. "Rationality, Morality, and Exit." *American Political Science Review* 86:418–431.

Vega-Redondo, Fernando. 2006. "Building up Social Capital in a Changing World." *Journal of Economic Dynamics and Control* 30:2305–2338.

Voss, Thomas. 2001. "Game-Theoretical Perspectives on the Emergence of Social Norms." In *Social Norms*, edited by Michael Hechter and Karl-Dieter Opp, pp. 105–136. New York: Russell Sage Foundation.

Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Watts, Duncan J. and Steven H. Strogatz. 1998. "Collective Dynamics of 'Small World' Networks." *Nature* 393:440–442.

Weesie, Jeroen. 1996. "Disciplining Via Exit and Voice." ISCORE Paper, Utrecht University.

Wilson, Roger. 1985. "Reputations in Games and Markets." In *Game-Theoretic Models of Bargaining*, edited by Alvin E. Roth, pp. 27–62. Cambridge: Cambridge University Press.

Yamagishi, Toshio, Nahoko Hayashi, and Nobuhito Jin. 1994. "Prisoner's Dilemma Networks: Selection Strategy Versus Action Strategy." In *Social Dilemmas and Cooperation*, edited by Ulrich Schulz, Wulf Albers, and Ulrich Müller, pp. 233–250. Berlin: Springer-Verlag.

Minor Change Is Not Enough: Analysis of Ebay's Reputation Model

Christoph Niemann, Stefan König, and Torsten Eymann

Chair of Information Systems Management
University of Bayreuth
95440 Bayreuth
Germany
christoph.niemann@uni-bayreuth.de
stefan.koenig@uni-bayreuth.de
torsten.eymann@uni-bayreuth.de

Abstract. One of the most popular eCommerce websites, Ebay, has changed its reputation model several times during the last months. Each change in the calculation of the reputation score affects the mean number of transactions the buyer is willing to participate in and the mean payoff for each buyer. While Ebay has never considered unrated transaction for its score, the meaning of neutral evaluations has changed throughout Ebay's lifetime. This paper analyzes the effects of different varieties of their reputation computation. A comparison shows that current model's aim is not to achieve high buyers' payoffs but to get a high number of transactions instead. In fact the current formula has a counter productive effect on the buyers payoffs. The analysis includes recommendations how to change the reputation system to avoid unwanted effects and repair the current weaknesses of the model.

1 Introduction

Since its start in 1997, the online auction website Ebay has used a reputation system to aid sellers as well as buyers to evaluate their potential transaction partners. However, the original system has been revised several times to address weaknesses in the model whenever they became apparent. The repair mechanism has frequently targeted the formula to calculate the so called reputation score of the transaction partner. Different incarnations of the formula each have different effects on the model's ability to make predictions on the honesty of the sellers in particular (because of the protocol that is commonly used on Ebay, a buyer has no opportunity to cheat on a seller).

This paper aims to analyze Ebay's current reputation model in terms of the effects it has on the number of transactions throughout the website. The second aim is to qualify the effects of the model on the buyers' payoffs. Buyers, of course, want to maximize their payoff using the site. If they continually end up with negative payoffs, their incentive to use the auction site decreases. Thus, for Ebay to remain a player in the market, it is essential to come up with a

reputation system that not only maximizes the number of transactions but the buyer's payoffs as well. We argue that the current reputation system does not achieve this two fold objective.

The structure of the paper is as follows: Section 2 explains the setting that Ebay uses from a game theoretic perspective as well as the reputation system in its current and former forms. The consequences of this setting are the focus of Sect. 3, which formulates two consequences and derives them theoretically. The paper concludes (Sect. 4) with recommendations on how to improve the reputation model to address the current weaknesses. Furthermore, avenues for future research are shown.

2 Environment

The environment of our analysis is the auction site Ebay[1]. Using this web site, customers are able to buy and sell items using a modified English Auction protocol. A seller sets up an auction by entering the specific characteristics of the item he wants to sell. The Ebay website assists the seller by providing a category system that the seller can use. Furthermore, the seller specifies a minimum bid that is necessary for the good to be sold as well as an end date for the auction.

A potential buyer can browse through the different items on the Ebay website or he can search for a specific item on the site. For each running auction, the buyer can set a maximum bid that he wants to place. A software agent supports the buyer by incrementing the actual bid by a minimal value until it is the only one bidder left (and thus the currently highest bidding agent) or until its maximum bid has been reached. The auction ends as soon as the end date has passed. The highest bidder at that time is awarded the item for his highest bid. As the software agents increase the bid only by a minimum value at each step, the auction protocol avoids the "winner's curse" [1] and approximates a "second price, sealed bid" auction [2].

To analyze Ebay's reputation system, the next section provides the foundations by describing a transaction on Ebay from a game theoretic perspective. This perspective includes the payoffs for both buyer and seller on Ebay. After the game's description, Sect. 2.2 describes the reputation model the Ebay uses.

2.1 The Ebay Trust Game

There are several research papers available [3–5] that analyze different aspects of various phases in the transaction phase model for transactions on eCommerce websites. From the perspective of game theory, the negotiation phase is the most interesting. Ockenfels has analyzed the reputation mechanism of the website "half.com"[2] and provides a description of the "Half.com trust game" [3]. The Ebay trust game shown in Fig. 1 is an adapted version of the Half.com trust game.

[1] www.ebay.com
[2] www.half.com

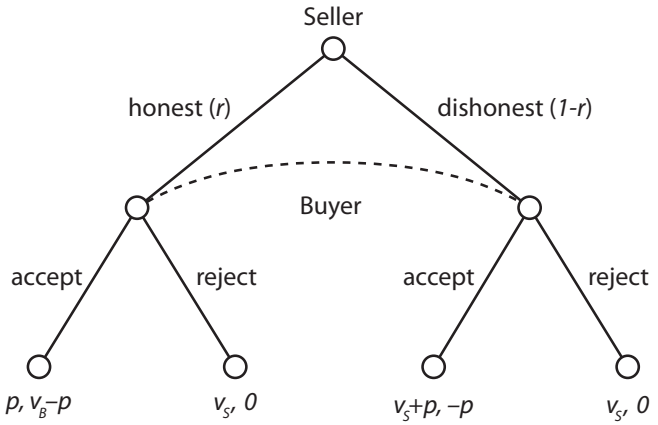


Fig. 1. The Ebay trust game (adapted from [3]): The dashed line indicates uncertainty about the buyer's position

After a buyer has won an auction, the commonly agreed upon protocol consists of four steps: Directly after the end of the auction, the buyer usually has to *pay* for the item. After the seller has received the payment, he *ships* the item. Once the item has been received by the buyer, the buyer has the option to *evaluate* the seller using Ebay's reputation system. Because of this model, only the seller may betray the buyer. If the seller receives the payment, he may opt not to ship the item. Thus, the seller can choose if he is honest, while the buyer has no accurate information about the particular seller. Figure 1 depicts this uncertainty with a dashed line. However, the buyer can use the reputation score ($r = [0, 1]$) to make an educated guess about the seller's trustworthiness.

In the game, firstly the seller chooses if he acts honestly. Secondly, the buyer has to choose, if he accepts the offer (by bidding in the auction) or if he rejects it by not participating in the auction. However, the buyer remains uncertain about the seller's decision (he does not know which node he is in). The payoffs for the different results are as follows:

Honest seller, buyer accepts: The seller parts with the item and receives the payment. His payoff is the price p. The buyer has to pay the price for the item and receives the item. His payoff is the price subtracted from his reservation price v_B of the object (manifest in his maximum bid) $v_B - p$.

Dishonest seller, buyer accepts: A betraying seller will keep the object after having received the payment. Thus his payoff is $v_S + p$. The buyer pays the price but does not receive anything in return. His payoff is $-p$.

Seller, buyer rejects: Since the buyer rejects the offer, no transaction takes place. The seller keeps the item valued with his reservation price v_S, the buyer does not receive any payoff. The seller's decision whether to act honestly does not make a difference in this case.

The single information that is available concerning the seller's reputation is the reputation score. A buyer on Ebay can interpret the seller's reputation score as conditional probability of shipment. Thus, a seller with a high reputation score probably ships the item, while a seller with a score close to zero probably does not. If the reputation score is considered as a probability of shipment, the expected payoff for a buyer is calculated as

$$EP(r,p) = r(v_B - p) + (1-r)(-p)\ . \tag{1}$$

After simplification this yields

$$EP(r,p) = rv_B - p\ . \tag{2}$$

A buyer who behaves rationally wants to maximize his expected payoff. If a buyer has a reservation price set for an item, the only way to achieve a high expected payoff is to choose sellers with a high reputation. Such sellers are likely to ship an item, which is the only outcome that can possibly lead to positive payoff for the buyer. A buyer who uses this strategy defines a suitable reputation threshold. The only sellers he trades with are the ones that have a reputation score equal or above the reputation threshold.

Another possible strategy of the buyer is to adapt his maximum bid depending on the seller's reputation: If the seller has a high reputation, the buyer is willing to place a bid that is close or equal to his reservation price v_B. If the seller's reputation is low, the buyer is willing to bid up to an amount $v_B - \varepsilon$ with a value for ε that depends on the seller's reputation. Figure 2 depicts the two strategies as maximum bid functions dependent on the seller's reputation.

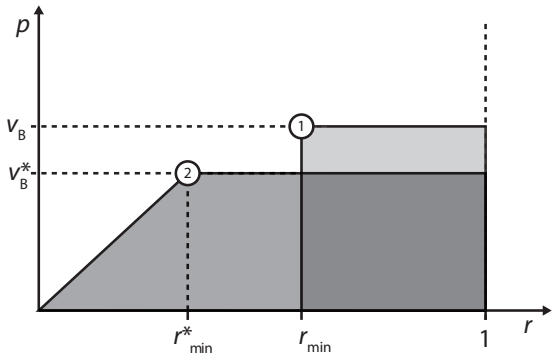


Fig. 2. The two strategies of the buyers: (1) is the threshold strategy, (2) depicts decreasing reservation prices with decreasing reputations. The shaded areas show all acceptable combinations of a seller's reputation and a buyer's price

Since both strategies are rational, the only assumption on the functions is that they are monotonically increasing. Strategy 1 is extremely focused on the

reputation value and excludes all sellers with a reputation value even minimally below this threshold. Strategy 2, on the other hand, provides a bit more flexibility: in addition to all sellers of strategy 1, it also considers sellers with a lower reputation level, but with a decreasing reservation price (deduction of risk). For the following analysis, the actual slope of the function is irrelevant as long as it remains monotonically increasing.

The next section describes Ebay's reputation model that forms the basis to evaluate a seller according to the history of his transactions.

2.2 Reputation Model

This subsection will consider the Ebay reputation model. Each seller and each buyer have to register on Ebay's website before selling or buying products. With the registration a new reputation account will be initialized. After buying a product from another Ebay participant, the buyer can evaluate the seller with a *positive*, *neutral* or *negative* evaluation. Each evaluation can be commented on by the evaluator. Evaluators sending negative ratings have to comment their rating anyway. In our further analysis of Ebay's feedback model, the comments are not considered anymore. After and only after the buyer has rated the seller, the seller is allowed to rate the buyer and his behavior during the transaction phases.

For the buyer's decision to participate in an auction, the feedback score of the seller seems to be the most important information source. The feedback score is computed by adding one point for a positive rating and subtracting one point for a negative rating. Neutral evaluations are ignored. For us, not the overall feedback score is of interest but instead, we will focus on the ratio of positive feedback. Especially the formula of this ratio is an important item that has been changed several times during the last months. The original reputation calculation was defined by

$$r_o = \frac{\text{positive}}{\text{positive} + \text{negative}} \,. \tag{3}$$

After changing the system in spring 2008, neutral evaluations are not been ignored any more, such that the new calculation is defined by

$$r_e = \frac{\text{positive}}{\text{positive} + \text{neutral} + \text{negative}} \,. \tag{4}$$

Neutral evaluations are now considered like negative ones in the formula. This change of the reputation calculation is based on the cognition that neutral evaluations are often given together with bad comments. [6]

The assumption that we follow in this paper is that even after this change the reputation does not represent the real behavior of the participant. Empirical studies [7, 8] show that many of the transactions remain unrated. In fact only about half of the transactions receive feedback. If they do, the overwhelming feedback is a positive score. So, we assume that unrated transactions are not

explicitly positive but have to be included in the computation. Buyers who have negative feelings about a particular transaction avoid to rate the seller, because they are afraid of a bad counter evaluation. This problem is often denoted as "Tit-for-Tat" artifact (see e. g. [9, 10]), or the crucial role of reciprocity [11]. If a buyer gives a bad evaluation in the central reputation system, he has to count on an equally bad counter evaluation, even if he has not made any mistake during the process. This fear of retaliation causes buyers to delay their negative feedback or refrain from providing it at all [12]. On the other hand, positive evaluations are provided very early, because buyers anticipate a reciprocated good evaluation. Although buyers cannot utilize a good reputation rating as such, they do have an incentive to collect good evaluations because Ebay uses the same evaluation rating regardless of the role (buyer or seller) that a participant is adopting.

The high dependency of the two (formerly independent) evaluations caused the Ebay management to change their system in a way such that only good evaluations can be answered by a seller's rating. The "tit-for-tat" problem, however, remains. The change just reduced the effects: Buyer's still hope for reciprocating good evaluations, only the seller's opportunity to retaliate negative evaluations is excluded from the set of possibilities.

This leads to our first hypothesis:

Hypothesis 1. *Ebay's reputation score r_e does not reflect the real behavior of the participant. Since unrated transactions are biased towards potential negative evaluations, they need to be considered in the reputation score.*

Three observations support the hypothesis. They are

1. the large number of unrated transactions,
2. the high number of positive evaluations contrasted with almost no neutral or bad evaluations, and
3. the problem of reciprocity.

To formalize the hypothesis we have to consider the combinations of four sets of reputational roles, the set of Beneficiaries (B), Memetic Agents (M), Evaluators (E) and last but not least Target Agents (T). In their work [11], the agent roles are defined as follows:

Set M is a group of agents that sends information to other agents. They act in a memetic way.
Set E are all agents which evaluate a certain target T.
Set T on the other side are the agents that are evaluated by E.
Set B is defined as a group of beneficiaries that benefit from evaluations performed by the evaluators (set E) about the target (set T) that can spread through the memetic agents (set M). The beneficiaries benefit from evaluations as they receive information about the degree to which the target conforms with the social norm. [11, p. 74 et seqq.]

The same authors are formulating hypotheses from the overlapping of these roles [11, p. 115 et seqq.]. The Ebay reputation mechanism does not differentiate

between the roles. The sellers can rate the buyers and vice versa. Additionally all participants can fulfil both roles, they can sell and buy products with the same identity, with the single exception of newcomers, who belong to set B only, because they have not engaged in any transactions yet. Practically all agents are playing all roles at the same time. In this case Conte and Paolucci state that this situation would be characterized by underrating and more positive evaluations. The first issue means that not all interactions are rated, the latter one that mostly the bad interactions remain unrated. This would lead to many positive feedbacks but missing negative ones. These hypotheses, derived from theoretical approaches, seem to provide exactly what we assume.

Although the reciprocating evaluations have not been studied empirically on the Ebay website itself, several authors assume their existence. For a broad overview on empirical studies we refer to [13]. In particular, Dellarocas et al. [14] analyze the Ebay setting as being prone the problem of overestimation of reputation. Masclet and Pénard [15] conducted an experiment that shows reciprocity: In the experiment, a setting that mirrored Ebay's reputation system resulted in a significantly lower ratio of negative to positive ratings than a setting that concealed the ratings until both parties had evaluated each other. The second setting avoids the opportunity to reciprocate a rating, which supports the hypothesis.

Assuming that hypothesis 1 holds, we propose a refined reputation calculation that provides a better predictor for the agents' behavior:

$$r_r = \frac{\text{positive}}{\text{positive} + \text{neutral} + \text{negative} + \text{unrated}} \,. \tag{5}$$

As a result, unrated transactions are treated in the same way as neutral and negative evaluations. Because the number of each evaluation type is greater or equal to zero, the following equation holds

$$r_o \geq r_e \geq r_r \tag{6}$$

for values of $x \geq 0$ for all components ($x \in \{\text{positive}, \text{neutral}, \text{negative}, \text{unrated}\}$) of the reputation values. This difference between the sellers' behavior (approximated by r_r) and their reputation (described by r_e) has severe consequences that are the focus of the next section.

3 Consequences

Based on the current reputation model, two hypotheses arise:

Hypothesis 2. *The advertised reputation r_e on the seller's reputation information page leads to a higher number of transactions than the reputation r_r that includes unrated transactions.*

Hypothesis 3. *The advertised reputation decreases the mean payoff for a buyer.*

The two hypotheses have an opposed effect on the expected payoff for a buyer that participates in a number of transactions. If hypothesis 2 holds, the expected payoff possibly increases, because the buyer participates in more transactions. If hypothesis 3 holds, the expected payoff potentially decreases. The following sections analyze the effects of both hypotheses.

Of course, the reputation value is not the only factor that influences a buyer's willingness to participate in an auction. Another one is the online presentation of an auction: A professional looking presentation can increase the average price by as much as 17.61% [16]. However, this paper focuses on the effects of the reputation system and excludes other influences.

3.1 H2: Ebay's Reputation Model Increases the Number of Transactions

In a setting with a number of sellers and a single buyer with either strategy this means that the number of possible transaction partners can only increase: If for either seller the assumption $r_r = r_e$ holds, a seller is considered based on the real reputation. If $r_r < r_e$ holds, buyers consider sellers on the basis of r_e which increases the seller's chance for a transaction. Whether the number of transactions increases with a the advertised reputation r_e in contrast to r_r depends on the strategy that a buyer uses. The following paragraphs analyze the two strategies in terms of their effect on the number of transactions.

Threshold Strategy. In case of strategy one (threshold strategy), a seller, who would be excluded based on his real reputation r_r, could be considered based on the advertised reputation r_e. Three cases for the relation of the threshold t and the reputation values r_e and r_r are possible.

$t < r_r < r_e$: If the threshold is below both reputation values, the seller is considered a potential seller based on r_r already. The advertised reputation r_e does not change that.

$r_r < t < r_e$: If the threshold is placed between the two reputation values, the seller would not have been considered based on r_r. However, he advertises r_e that is greater than t. Thus, the buyer does not exclude the seller based on his reputation; the number of transactions potentially increases.

$r_r < r_e < t$: If the threshold is greater than both r_r and r_e, the seller is neither considered based on r_r nor based on r_e.

Therefore, if the advertised reputation is increased in relation to the real reputation, the number of total transactions can only increase. A seller whose advertised reputation jumps to a value above the threshold equals an additional transaction. As there is no possibility to *reduce* the advertised reputation, the number of total transactions cannot decrease.

Decreasing Reservation Prices. A buyer that uses strategy two (decreasing maximum bid with decreasing reputation), participates in the same number of transactions regardless of the advertised or real reputation of the seller. However, as he modifies the maximum bid according to the seller's reputation, a higher advertised reputation r_e leads to higher bids than the buyer would have set with a seller revealing r_r. As a consequence, the expected payoff decreases, because the value of p in (2) is closer to the buyer's true reservation price v_B.

Ebay collects a transaction fee for each successful transaction that is usually paid by the seller. This transaction fee provides an incentive for Ebay to maximize the number of transactions because Ebay's profit increases with the number of successful auctions. This incentive is not necessarily compatible with the goals of buyers on Ebay who want to maximize their payoffs.

3.2 H3: Ebay's Reputation Model Decreases the Buyer's Payoff

After showing that a reputation model like the current one at the Ebay website increases the number of transactions, this subsection will show that the Ebay model decreases also the buyer's payoff. Like denoted in (2), the expected payoff of a buyer is $rv_B - p$. The assumption of rationally acting buyers implies that a buyer will participate in an auction if and only if the expected outcome is positive. Thus, he wants to maximize the payoff under the condition

$$EP(r, p) = rv_B - p \geq 0 \; . \tag{7}$$

The result of the maximization of the expected payoff depends on an assumption on the type of market that Ebay represents. If one assumes Ebay to be a perfect market the results differ from the assumption of Ebay being an imperfect market. The following paragraphs analyze the outcome for the two market types.

Ebay is a Perfect Market. In a perfect market the reservation price v_B of the buyers will tend to converge on the market price of the item. Brynjolfsson et al. for example conclude [17], based on empirical findings, that online auctions nearly fulfill the conditions of a perfect market. Thus, in a perfect market, one can assume $p = v_B$. Simplifying (7) under this assumption leads to

$$r \geq 1 \; . \tag{8}$$

As r is always in between 0 and 1, this means that rational buyers are only willing to buy products if they can be absolutely sure that the seller will deliver the product, i. e. the reputation value equals 1. If the reputation is less than 1, the expected payoff for the buyer is negative. Hence, buyers on an (assumed) perfect market can avoid negative payoffs by participating in auctions only if a seller above suspicion of betraying.

Ebay is an Imperfect Market. If Ebay does not represent a perfect market, the buyer's reservation price is greater than the market price: $v_B > p$. That means the buyer does not need to bid the exact market price; products are sold below it. Like in the perfect market case, the expected payoff for a buyer must be positive to have a rational buyer remain in the market. Starting from (7) again, simplification leads to

$$r \geq \frac{p}{v_B} \tag{9}$$

for $0 \leq r \leq 1$ and $v_B, p \geq 0$. With $v_B > p$, the buyer's reputation threshold is less than 1. If a buyer adjusts the maximum bid to account for the reputation of a seller, he can use the advertised reputation r_e only. Since a buyer does not want to fall for the winner's curse, the maximum bid he can use is $r_e v_B$ (his reservation price discounted with the seller's reputation). The rational buyer would set the maximum bid such that the expected payoff remains zero. However, as $r_r \leq r_e$ (the real reputation is less or equal to the advertised reputation), the modification of the maximum bid proves to be too small in the worst case ($r_r < r_e$): The buyer should have discounted for the real reputation r_r but has used r_e instead. Such a setting leads to a negative payoff for the buyer. In the best case ($r_r = r_e$) the payoff remains at zero, because the discount rate is just at the right value. However, with the large number of unrated transactions found in reality (and the corresponding difference in r_r and r_e), the chances to keep a payoff of zero diminish.

To conclude, if Ebay is assumed to be a perfect market, buyers can reach a payoff of zero at best. If Ebay is seen as an imperfect market, the buyer's payoff will be negative in the majority of cases. The extent of the negative payoff depends on the difference in the reputation values r_r and r_e.

4 Conclusion and Future Work

In conclusion, we found out that the current Ebay reputation system does not work very well in the perspective of the buyers. If hypothesis 1 is assumed to be true, the advertised reputation score is too high. Basing the decision for a maximum bid on the advertised reputation, the payoff will turn out to be negative, because the actual behavior includes a higher probability of deception. Our suggested reputation model highlights the identified shortcomings, but does not provide a useful model in practice, because it introduces new problems, such as dysfunctional incentives, for example the inducement not to rate in case of a negative outcome of a transaction.

The shortcomings of the reputation system are based in most of the cases on the effects of the "Tit-for-Tat" problem; they occur because the set of evaluators is nearly the same as the set of beneficiaries. Thus, in Ebay's reputation model no reputational roles are considered, which means that the group of targets overlaps with the other roles.

The suggestion to introduce different reputational roles [11] could solve the issue of retaliation. Another possibility is to abandon the seller's opportunity to

rate the buyers. Both strategies disentangle the reputational groups and separate the group of targets from the group of evaluators. That way the system disables the sellers from reciprocating (both positive and negative) feedbacks.

However, in reality a drastic change like that might be impossible to implement. Therefore an approach with incremental changes seems to be more promising not least because of the preservation of existing information. Furthermore the suggested mode of introduction would be to provide additional feedback opportunities instead of substituting the system [18], which is more or less what Ebay did. In particular, the change to include neutral evaluations in the score has been a step compliant with our hypotheses, however, the revocation in August 2008 emphasises the problem again.

This article highlighted some shortcomings in Ebay's reputation system. The consequences are conditional on the assumption that the advertised reputation does not predict the seller's behavior but rather overestimates it in terms of honesty. This assumption has been studied in experimental settings as well as from a theoretical perspective. However, empirical studies with a real Ebay sample could substantiate the hypothesis.

This paper *qualified* the effects of the difference in reputation scores. We plan to continue to work on the paper and *quantify* the effects as well. Only after it has been shown which effect (if one at all) dominates the other, an evaluation on the consequences becomes possible.

A third avenue for future work is the possible dependency of potential evaluations on the item to be sold. Intuition suggests that the probability to rate truthfully would increase with the value of the good (because of the potentially much higher utility gain or loss). Such a dependency could provide ample possibilities for research.

Furthermore, our model is still oversimplified in the current state. Future work might include a stepwise adaption of the model to the real Ebay reputation system (e.g. consideration of the time dependency of ratings' visibility or the textual comments on transactions).

References

1. Bajari, P., Hortacsu, A.: The winner's curse, reserve prices and endogenous entry: Empirical insights from ebay auctions. RAND Journal of Economics **34**(2) (2003) 329–355
2. Wurman, P.R., Wellman, M.P., Walsh, W.E.: The michigan internet auctionbot: A configurable auction server for human and software agents. In: Proceedings of the Second International Conference on Autonomous Agents (Agents-98). (1998)
3. Ockenfels, A.: Reputationsmechanismen auf Internet-Marktplattformen – Theorie und Empirie. Zeitschrift für Betriebswirtschaft **73**(3) (2003) 295–315
4. Kumar, M., Feldman, S.I.: Business negotiations on the internet. In: INET98 – Conference of the Internet Society. (1998)
5. Bakos, Y.: The emerging role of electronic marketplaces on the internet. Communications of the ACM **41**(8) (1998) 35–42

6. Botsch, J., Luckner, S.: Empirische Analyse von Bewertungskommentaren des Reputationssystems von eBay. In: Multikonferenz Wirtschaftsinformatik 2008. (2008)
7. Resnick, P., Zeckhauser, R.: Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. the economics of the internet and e-commerce. In Baye, M.R., ed.: Advances in Applied Microeconomics. Volume 11., Amsterdam, Elsevier Science (2002) 127–157
8. Sun, C.H., Hsu, M.F.: The determinants of price in online auctions: More evidence from quantile regression. Technical Report 07-18, University of Wollongong (2007)
9. Axelrod, R.: The Evolution of Cooperation. Basic Books (1984)
10. Axelrod, R.: The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration. (1997)
11. Conte, R., Paolucci, M.: Reputation in Artificial Societies – Social Beliefs for Social Order. (2002)
12. Klein, T.J., Lambertz, C., Spagnolo, G., Stahl, K.O.: Last minute feedback. Working paper, University of Mannheim (2007)
13. Dellarocas, C.: The digitization of word of mouth: Promise and challenges of online reputation mechanisms. Management Science **49** (2003) 1407–1424
14. Dellarocas, C., Dini, F., Spagnolo, G.: Designing reputation (feedback) mechanisms. In: Handbook of Procurement. Cambridge University Press (2006)
15. Masclet, D., Pénard, T.: Is the ebay feedback system really efficient? – an experimental study. Working paper WP 2008–03, University of Caen (2008)
16. Gregg, D.G., Walczak, S.: Dressing your online auction business for success: An experiment comparing two ebay businesses. MIS Quaterly **32**(3) (2008) 653–670
17. Brynjolfsson, E., Dick, A.A., Smith, M.D.: A nearly perfect market. SSRN eLibrary (2008)
18. Bolton, G., Greiner, B., Ockenfels, A.: Engineering trust: Strategic behavior and the production of reputation information. (2008)

A Reputation System for Uncertain Assertions

Mark Kramer, Arnon Rosenthal
The MITRE Corporation, 202 Burlington Road
Bedford, MA, USA
{mkramer, arnie}@mitre.org

Abstract. We investigate reputation systems that rate the performance of analysts who make uncertain assertions (claims accompanied by estimated probabilities). Accuracy metrics (based on the fraction correct) are fair only if all analysts handle identical or statistically similar cases. Furthermore, accuracy metrics discourage analysts from offering predictions on difficult-to-predict events. Because of these difficulties, we develop a class of performance scoring functions that are maximized when the analyst provides accurate probabilities, especially when these probabilities differ from the norm. Under these metrics, the disincentives to forecast low-probability events is removed and analysts are rated fairly, independent of the base event probabilities of the cases they consider. Reputation systems built around these metrics can support productivity management and increase manipulation resistance when information providers are not trustworthy. An application to citizen event reporting is presented.

Keywords: uncertainty, reputation, scoring function, citizen event reporting, counter insurgency

1 Introduction

Many analysis tasks, such as economic forecasting, criminal investigation, and medical applications, produce statements and predictions that involve significant uncertainty. In these cases, probabilistic statements are frequently more useful than simple Boolean predictions. For example, if we are planning an outdoor activity, generally we prefer predictions in the form "it will rain tomorrow with 30% probability", instead of getting an unqualified prediction.

Now suppose there are multiple information sources (human or otherwise) providing probabilistic predictions. In planning courses of action, it would be useful to have meta-knowledge of the trustworthiness and likely novelty of the predictions. For example, a military commander who relies on knowledge of probable enemy responses would clearly benefit from knowing the skill of analysts providing the predictions.

Prediction skill differs for many reasons. Some analysts possess information unavailable to others, such as an equity analyst able to communicate with key company insiders. Others may have superior subject matter expertise or longer, richer experience. They may use different techniques, and apply different levels of skill and judgment. Moreover, information providers can be honest, manipulative, or

malicious. The chance of malicious behavior can be minimized with carefully screening of information providers, but in some contexts, this is not possible. As discussed further in Section 4, citizen event reporting (CER) leverages the eyes and ears of a large population of "citizen sensors" to increase the amount of information available to decision makers. When deployed in an environment that includes a hostile subpopulation or rival clans, some of the tips gathered by CER may be aimed at deceiving decision makers, motivated by the desire to lure first responders into an ambush or induce them to attack rivals. In such a case, it is advantageous to track the history of reports obtained from the CER system, to help determine the trustworthiness of the reporters.

Providing meta-knowledge about information providers is the job of a reputation system. A reputation system can track the success of information providers over time, and provide rankings, feedback, and other information useful to the participants (information providers, consumers, or both). In this paper, we assume a centralized reputation system that possesses global information about all analysts and their prediction history. In this context, the simplest reputation system would track *accuracy*, the fraction of correct answers provided by each analyst. Unfortunately, accuracy provides a fair comparison only if all analysts consider the same or equivalent set of cases, in terms of their prior likelihood and intrinsic predictability. If analysts exercise the freedom to choose when they make predictions, then to maximize their accuracy, they will avoid making assertions about low-probability events, focusing instead on "sure things". Obviously-correct assertions (the sun will rise tomorrow) have low value to information consumers. In general, the use of accuracy to rank analysts creates a mismatch between the needs of information consumers and information providers.

The goal of this paper is to propose a reputation system for uncertain assertions. When assertions include an estimated probability, it is possible create a system of rewards that does not skew the attention of analysts towards high-probability events. Rather than rewarding accuracy itself, our ranking system rewards both novelty, in terms of departure of a case from the norm (expressed in terms of the consumer's prior), and accurate probability estimates. To this end, we first examine classes of scoring functions that are aware of prediction probabilities, and derive measures of producer accuracy and productivity. Next, we address how reputation scores can be used, and explore assumptions about provider motivations and work processes. For user organizations, the framework and analysis help identify questions one must ask in setting up a reputation system. For reputation researchers, the framework can help in capturing assumptions and comparing with others' results, and in identifying gaps in our knowledge. We emphasize metrics that require relatively little software and administrative labor to implement – though sometimes less accurate, they seem far more likely to be implemented.

2 Reputation Metrics

2.1 Preliminary definitions

We now define the constituents of the model and the notation to be used in the rest of the paper. For the purposes of this paper, we consider predictions consisting of a Boolean assertion about the world (A), and the provider's estimated probability ($0 \leq Q \leq 1$). An *uncertain assertion* is a pair (A, Q). Examples include:

- {it will rain in Boston tomorrow, 0.7}
- {patient Z will survive the proposed surgery, 0.9}
- {there is a roadside bomb on next route segment, 0.01}

Other types of predictions, such as quantitative, multi-valued predictions, or interval probabilities, are not considered here, but some discussion is available in [3].

An assertion may be about the past, present, or future, but reputation points will be assessed only when its truth (or falsity) becomes known. *Assertions for which we learn ground truth drive the assignment of reputation.* Ground truth is frequently available in areas such as weather, elections, and sports, and less often in fields such as medicine. We now introduce several definitions:

- *Event probability* (P): Each assertion is Boolean and may be decided at some future time. However, while the outcome remains unknown, there is uncertainty in the outcome[1]. We do not expect administrators to know or estimate P; it appears only in our mathematical analysis.
- *Prediction* (Q): Denotes the provider's estimate of P, a probability between 0 and 1.
- The *prior* (P^*) is the consumer's probability estimate for A, occurring before the provider information is taken into account. If the prior is not known, it defaults to the uninformative prior (usually 0.5).
- *Utility* (U) measures the value of learning the truth or falsity of A. When determining reputation, scores relative to assertion A can be weighted by the utility of A.
- *Scoring (payoff) functions.* We define two mathematically-related functions, $f(Q, P^*)$ and $g(Q, P^*)$, representing the analyst's reward if A does or does not occur, respectively. Since a strategy for optimizing expected score across multiple independent assertions will attempt to maximize the score separately for each assertion, we can consider scoring a single assertion without loss of generality.

We consider it essential to minimize the administrative effort, i.e., the effort to ascertain scoring functions, priors, utility, and ground truth. If a reputation system

[1] For the interpretation of probability for non-repeated event, see [1]. For our purposes, P denotes the fraction of instances like the current situation in which A holds.

requested additional information about each assertion, most providers and managers would probably refuse, or provide perfunctory estimates. For that reason, the default treatment is to use the same scoring function for all assertions, and equal weighting for utility. Still, if a large number of assertions have the same properties (e.g., weather assertions for different days), it may be feasible to elicit a few numbers (prior probability, utility, perhaps choice among loss functions) for different classes of assertions, to be applied to each instance.

2.2 Reputation Scoring Functions

Various scoring functions have already been developed with the objective of rating forecasting skill and eliciting truthful personal beliefs [2, 3, 4, 6]. Much of the work originated in the field of weather forecasting. Most commonly used is the Brier score [2], defined as:

$$f(Q) = -(1 - Q)^2 \qquad (1)$$
$$g(Q) = -Q^2$$

The Brier score is the squared error based on the difference between the predicted (Q) and actual (0 or 1) probabilities (the negative sign makes higher scores better). It is independent of the prior probability, which is a drawback, because a provider's assertion is useful only if it is different than the consumer's prior. We believe a reward should be given only if the consumer learns something new (and correct) from the information provider. For example, the assertion {the sun will rise tomorrow, 1.0} has little surprise or impact, while {terrorists will attack Mumbai, 0.10} might be quite novel and useful. A higher score should be awarded for the latter assertion, if proven true, even though the asserted confidence is lower, because the difference between the prior and the prediction is larger.

We have identified several requirements for a probability-aware payoff function, as follows:

R1. **Monotonicity**. We require that $f(Q, P^*)$ increase monotonically and $g(Q, P^*)$ decrease monotonically as a function of Q. This assures that if the event occurs, larger estimated probabilities earn larger rewards. Conversely, assertions that are confident but wrong receive larger penalties than wrong answers that are explicitly declared uncertain.

R2. **No Reward for Prior**. As discussed above, a provider's assertion is useful only if it is more informative than the consumer's prior. Accordingly, we require that $f(P^*, P^*) = g(P^*, P^*) = 0$.

R3. **Expected Value**. For a Boolean assertion, the expected value of a prediction is $E(Q, P, P^*) = P\, f(Q, P^*) + (1 - P)\, g(Q, P^*)$. So each analyst is motivated provide a prediction as close to the true probability as possible, we require that $E(Q, P, P^*) < E(P, P, P^*)$ for each $Q \neq P$. A scoring rule with this property is said to be *strictly proper* [3]. This requirement assures that, over the long run, the perfect information provider who picks Q = P in each situation will earn the maximum possible score.

R4. **Hidden Knowledge**. The scoring functions cannot require knowledge of P, since this probability is assumed to be hidden from all parties.
R5. **Boundedness**. No assertion should earn an unbounded payoff (either positive or negative), because this would make accumulation of scores over multiple trials problematic.
R6. **Symmetry under Negation**. Asserting A with certainty Q is the same as asserting ~A with certainty 1-Q. Our payoffs should be indifferent to the logical "direction" of the assertion. Therefore, we require that $f(Q, P^*) = g(1–Q, 1–P^*)$ and $g(Q, P^*) = f(1–Q, 1–P^*)$.

If we assume f and g are differentiable functions, then R3 implies that $dE/dQ = 0$ at the point where $Q = P$. Therefore, strictly proper scoring functions must satisfy the differential equation:

$$P\, df/dQ + (1 – P)\, dg/dQ = 0 \text{ at } Q = P \qquad (2)$$

Furthermore, to satisfy R2, we can assume that scores depend only on the difference between the estimated probability and the prior, denoted $\Delta = Q – P^*$ (this is sufficient but not necessary). Under that assumption, $f(Q, P^*) = f(\Delta)$ and $f(\Delta=0) = 0$. In the following, we give examples of strictly proper scoring functions that satisfy these constraints.

Log/Linear Payoff. By means of example, suppose $f(\Delta) = \Delta$. This choice trivially satisfies the monotonicity condition (R1), and produces a zero reward for guessing the prior (requirement R2). To satisfy the expected value condition (R3), we solve Eq. (2). This simplifies to $P + (1 – P)\, dg/dQ = 0$ at $Q = P$, or $dg/dQ = –Q/(1 – Q)$. It follows that $g = Q + \ln(1–Q) + c$. Applying the boundary condition that $g(P^*, P^*) = 0$ (again R2) we obtain the following scoring functions:

$$f = Q – P^* \qquad (3)$$
$$g = Q – P^* + \ln[(1 – Q)/(1 – P^*)]$$

These functions satisfy requirements R1 through R4. However, g is unbounded, approaching $-\infty$ as $Q \rightarrow 1$, violating requirement R5 (see Fig. 1). Additionally, the symmetry condition (R6) is not met. Two information providers who respectively assert A with certainty Q, and ~A with certainty $1 – Q$, would receive different rewards. Finally, the expected value curves are essentially flat over large ranges, implying the reward function will not effectively discriminate among analysts, provided they stay away from near-certain predictions. These disadvantages eliminate log/linear payoff from further consideration.

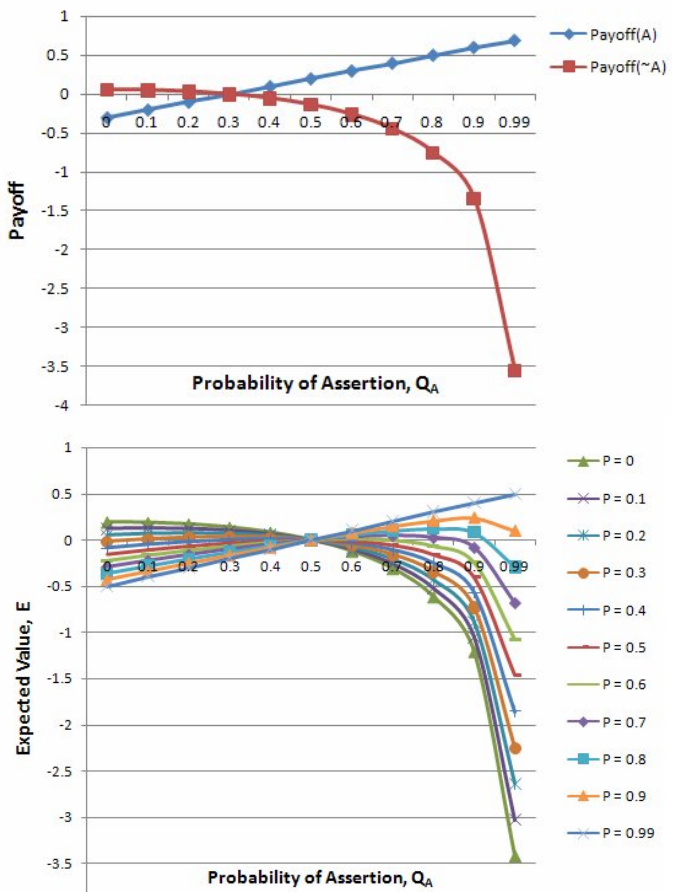


Fig. 1. The linear/log payoff function (Eq. 3), showing the payoff functions (top) and expected values (bottom). The example is for $P^* = 0.3$.

Quadratic Payoff. Assume f (Δ) has the form $a_0\Delta^2 + a_1\Delta + a_2$. We know $a_0 < 0$ because E must be concave downward for all values of Q. We can choose $a_0 = -1$ to specify the scoring function to within a multiplicative constant (which may be chosen as a function of the prior). In addition, $a_2 = 0$ (based on R2) so that f (Δ) = 0. Solving Eq.2 and applying boundary conditions (math omitted):

$$f = 2(1-P^*)(Q-P^*) - (Q-P^*)^2 \qquad (4)$$
$$g = -2P^*(Q-P^*) - (Q-P^*)^2$$

These functions are depicted in Fig. 2. Unlike the linear case, the payoff functions remain bounded, meet the symmetry condition, and do not have large flat regions. As an example of symmetry, f(0.8, 0.3) = g(0.2, 0.7) = 0.45.

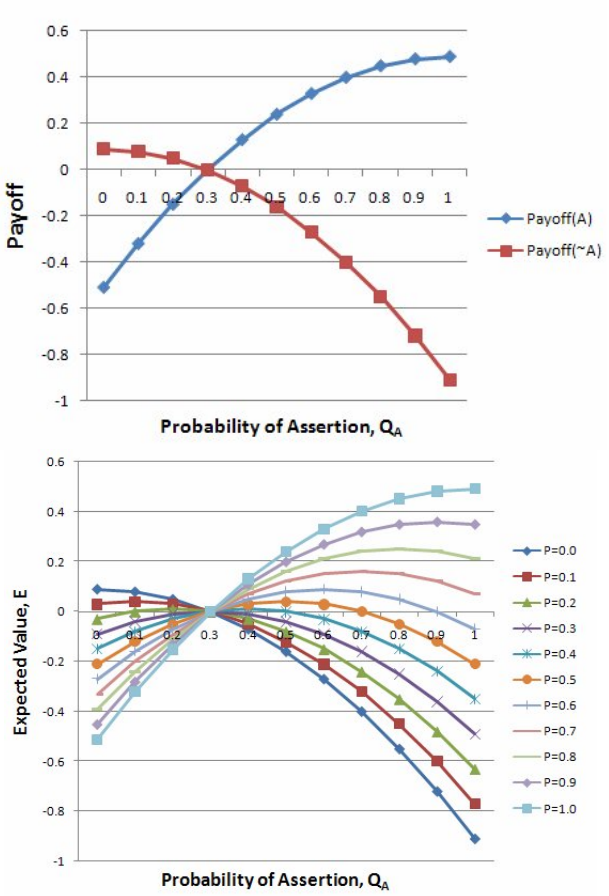


Fig. 2. The quadratic payoff function (Eq. 4), showing the payoff functions (top) and expected values (bottom). The example is for $P^* = 0.3$.

Other Payoff Functions. Among differentiable functions, we can also show that binomials in the form:

$$f = -\Delta^c + (1- P^*)(c/(c-1))\Delta^{c-1} \qquad (5)$$
$$g = -\Delta^c - P^*(c/(c-1))\Delta^{c-1}$$

generate admissible solutions for powers of $c \geq 2$ and integral. An open problem is to explore the usefulness of c fractionally greater than 1, a range that offers greater discrimination for Q near 1. The treatment will need care to avoid generating imaginary roots for negative delta. One might also explore multiplicative formulations, rewarding for the fractional change between P* and 1. However, the quadratic payoff satisfies all our conditions, and is certainly the simplest pair of functions to do so.

2.3 Analogy: Weighted Coin Tosses

In this section, we present an analogy to the current problem of predicting the probability of future events. Suppose we have a bag of coins confiscated from dishonest gamblers. Each coin may be weighted, so the chance of heads or tails is not 0.5. However, there is no reason to assume there are more coins weighted towards heads than tails. Therefore, the overall prior probability of heads is 0.5. We randomly select a coin from the bag (this represents a situation requiring analysis), and allow multiple analysts to physically examine the coin, without tossing it. Each analyst then predicts the probability of tossing heads with the coin. The coin is then tossed (producing ground truth). If the toss is heads, each analyst is paid off according to f, and if the toss is tails, each analyst is paid according to g.

Consider the following analysts:

1. *Probabilistically perfect analyst.* Always predicts the probability of A (heads) exactly. Note that "perfect" denotes accurate probability assessments, as opposed to the (impossible) clairvoyant analyst, who accurately predicts the outcome of each coin toss.
2. *Random analyst.* Produces uniform random guesses between 0 and 1.
3. *Biased analyst.* Over- or underestimates the chance of heads by a fixed amount, except where such a prediction would exceed 1 or go below 0.
4. *Noisy analyst.* Each probability estimate is off by random number drawn from a normal distribution with zero mean and given standard deviation (bounded between 0 and 1).
5. *Prior analyst.* This analyst always predicts the prior probability.

Figure 3 shows the results of 500 trials using the quadratic scoring function (Eq. 3). As expected, the perfect analyst outscores the other analysts in the coin-assessment task. The noisy analyst, shown for standard deviation 0.25, is the second best. The biased analyst, who in this case overestimates the probability of heads consistently by 0.25, is next. The analyst who picks the prior probability (0.5 in this case) earns zero, and the random analyst loses points at about the same rate as the perfect analyst earns points.

When the same experiment is carried out with the linear/log payoff function (not shown), the biased and noisy analysts both eventually make certain predictions (probability 0.0 or 1.0), which subsequently turn out to be false, causing them to earn $-\infty$ points, and fall off the chart.

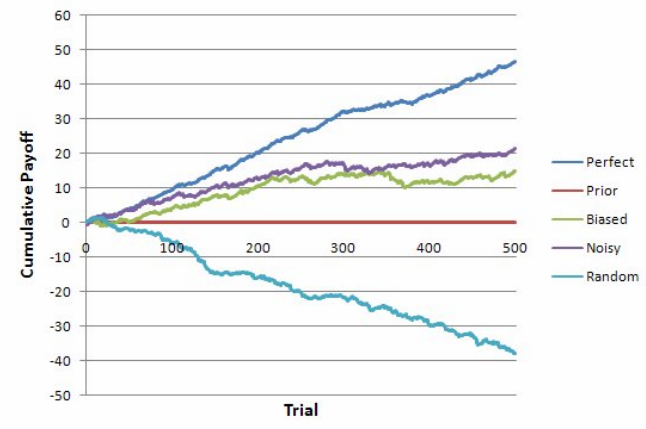


Fig. 3. Result of the weighted coin analogy for five types of analysts. In the case shown, the biased analyst consistently overestimates the probability of heads by 0.25, and the errors of the noisy analyst follows a normal distribution with standard deviation of 0.25.

Figure 4 shows the average payoff as a function of the magnitude of analyst error (bias for the biased analyst, standard deviation for the noisy analyst). In this chart, the average payoff for the perfect analyst (0.083) has been normalized to 1. Both the biased and noisy analysts perform worse than the random analyst for large biases or standard deviations because their probability predictions tend to the extremes. For example, the analyst suffering a large bias will consistently call heads (or tails) with probability 1, which is much worse than picking a random probability. The noisy analyst outperforms the biased analyst, because even in the extreme, some probability predictions will be between 0 and 1.

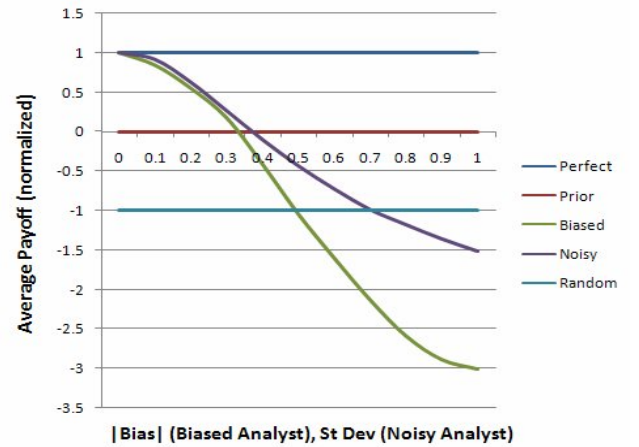


Fig. 4. Average reputation points earned per trial, for each type of analyst, as a function of average error (normalized to perfect analyst = 1).

2.4 Normalization and Analyst Comparison

Given scores for individual assertions, reputations can be developed by aggregating scores. The most obvious methods are summing and averaging. Averaging compensates for the difference in the total number of assertions scored by each analyst. However, averaging still does not ensure a fair comparison between information providers. Consider information utility. Information consumers might consider some assertions more important than others, and (formally or informally) assign a utility score to each assertion. In this case, the accumulated reputation score should derive from a weighted sum involving the product of the raw score and the utility of each assertion. The utility might be assigned by the consumer at the personal or enterprise level, the former leading to a personalized set of most trusted information providers. To reduce workload of manual assignment of utility, we suggest doing assignments to classes of assertions (e.g., probability that a patient has a certain disease). A default (utility=1) is also provided, lest additional administration discourage participation of the reputation system.

The other desirable correction involves differences in the priors between different analysts. The admissible scoring formulae are based on Δ, the difference between the asserted probability and the prior. In this way, we reward the analysts' unique contribution (novelty) as well as accuracy. To help understand the impact of priors, consider two doctors who analyze MRIs to diagnose a certain disease. The first doctor examines all MRIs. His prior reflects the incidence of the disease in the overall patient population. The second doctor is consulted for a second opinion only when the first doctor suspects the presence of the disease. Her prior reflects the incidence of disease in the subpopulation identified by the first doctor. It is easier for the first doctor to earn reputation points, because his observations change the probability of disease to a greater extent than the doctor offering a second opinion. These two different priors reflect different opportunities to earn reputation points, and should be accounted for when comparing and ranking the doctors.

To remove the impact of unequal priors, we can normalize by the size of scoring opportunity using either of the following two maximum obtainable scores:

- *Factually perfect* represents the score obtained when someone receives the maximum score at every opportunity. We obtain this standard by assuming $Q = 1$ when the event occurs and $Q = 0$ when the event does not occur.
- *Probabilistically perfect* represents the expected value of the score if the analyst chooses $Q = P$ at every opportunity. Since the scoring formulae explicitly maximize the expected value of the score for a probabilistically perfect analyst, this standard represents the best possible *obtainable* performance by any analyst.

To take the concept of probabilistic perfection slightly further, we can calculate the maximum expected value E_{max} under the assumption that $Q = P$, for the quadratic scoring function:

$$E_{max}(P, P^*) = E(P, P, P^*) = P\, f(P, P^*) + (1 - P)\, g(P, P^*)$$

$$= (P - P^*)^2 \qquad (5)$$

Based on Eq. 5, the *opportunity* to gain reputation points is proportional to the square of the difference between the prior and the actual event probability. In fact, if the consumer possesses an accurate prior for each event, the long-run scoring opportunity for any analyst is zero, since the analyst cannot tell the consumer anything new (although luck might prevail in the short run).

To apply Eq. (5) we must know P, which can only be estimated from relevant historical information, which may not exist. Alternatively, one may use a proxy for P, such as a group judgment. Using group judgment as a gold standard introduces a host of potential problems, such as unfairly downgrading independent-thinking analysts. Using the factually perfect score (the clairvoyant analyst) as the normalizing factor is a possibility, because it can be calculated using ground truth, but the effectiveness of approach is an open problem.

The last issue here is extrapolating to non-scored assertions. As we have discussed previously, assertions without ground truth are not scored. But given Q, the analyst will receive only one of two possible scores, f or g, the former with probability P and the latter with probability (1–P). Given P, we can determine the expected value of the score without ground truth. However, as in the case of normalizing for unequal opportunity, we are confronted with the unknown event probability. Again, one may use some proxy for P, such as group opinion, as long as one is aware of the inevitable hazards.

3 Other Considerations

3.1 Provider Behaviors

As mentioned earlier, information providers fall into three classes of behavior: honest, manipulative, and malicious.

Honest behavior has the provider doing his or her best, regardless of the reputation system. Obliviously honest behavior is likely among dedicated employees or in situations where there is no benefit to manipulation. A provider who is honest will attempt to estimate probabilities that match the true event probabilities.

Manipulative behavior aims to inflate reputation scores. The reputation system may be linked to rewards, privileges, and prestige, and thus a manipulative user may wish to accumulate undeserved rewards. As a side effect, but not as a goal, manipulators may deceive consumers about event probabilities. Manipulation can occur in two ways:

- *The choice of questions to answer.* The provider might attempt to "cherry pick" opportunities with high utility, e.g. a major crime, even if they have no special knowledge or qualification relative to the case. If evaluated by average accuracy, they may only choose cases where they have high certainty. Conversely, if scores are not weighted by utility, an analyst can manipulate the system by reporting on many obvious or uninteresting phenomena. Within an organization, this may be

controlled by management oversight, for example, by penalizing time-wasting or by randomly assigning cases to analysts.

- *The probabilities.* A manipulator who seeks to optimize expected score has no incentive to mis-estimate probabilities, since in the long run, the correct probability gives the highest expected score. However, a manipulator might over- or under-estimate probabilities in the short run, seeking the "big win" or seeking to avoid a "loss". It is thus desirable for management to reward long-term success, rather than rewarding occasional big wins or punishing mistakes.

An alternative route to combat manipulation is to keep providers unaware of the scoring system. This may be unethical within an enterprise, but it is certainly feasible for rating external providers, such in citizen event reporting or rating stock market pundits on the web.

Malicious behavior seeks to fool consumers, i.e., to convince consumers to believe and act on an incorrect probability. Methods for discouraging this behavior are well known. Casual attackers who seek instant gratification but will not invest much effort can be discouraged by requiring them prove identity, or requiring a certain number of previous postings that have been determined to be accurate before accepting their recommendations. The determined attacker is more difficult. One way to discourage sustained attack is to require that analysts provide useful information whose total utility is greater than the dis-utility of their deception. Thus, in return for possibly selling the lie, the attacker must do considerable work for the benefit of the organization. If they do not know the threshold for acceptance, this adds to their difficulty.

3.2 Applications within the Enterprise

In an enterprise, management might use scores in several ways in their relationship with employees, or with other sources that are recruited to provide information. In this section, we briefly discuss the uses for the reputation metrics by management that go beyond rating information providers.

Training. Management can teach providers how to remedy their identified weaknesses. If accuracy seems high (relatively) but productivity (quantity) is low, one might consider working more quickly. If accuracy seems low, one may need to examine the analyst's techniques and improve subject-matter expertise. If utility weights are available, these can help the analyst focus on important cases. Bias estimates may help providers adjust for undue optimism or pessimism.

Assigning workload. Management may be responsible for giving each provider suitable tasks. Here, both accuracy and productivity is important. For example, a provider who has been accurate on relatively unimportant tasks might be given more important ones. Less critical tasks might be assigned to a provider who is moderately accurate but very fast. Finally, a provider who has shown ability to select high-payoff tasks can be given more freedom.

Judging and improving accuracy. Some providers may consistently overestimate probability of their assertions, while others may underestimate. Bias statistics attempt to estimate these tendencies, and can also be used to revise probability estimates.

4. Example: Citizen Event Reporting

In this section, we demonstrate the reputation system in the context of citizen event reporting (CER). CER involves using citizens to bolster information collection. This approach has most commonly been applied to crime-fighting efforts, but recently CER has been considered for asymmetric warfare [5]. The conflicts in Iraq and Afghanistan have highlighted the need to gather information known primarily by the local populace. Of particular concern is detection of improvised explosive devices (IEDs), bomb-making facilities, and identification of militant insurgents and terrorists. The challenge is to make CER work in hostile environments, where enemies may contribute false reports in an attempt to "game" the system, to lure first responders into ambushes, create decoys, or induce the authorities to target third parties.

In our example, we utilize a modified version of the agent-based simulation reported in [5]. In this simulation, there is one type of event to be reported, which we call a fire, but could represent sighting of a suspicious person, a weapons cache, IED, etc. Simulated citizens traverse the environment and can report an event if they come within a certain distance of it and "see" it. Reports represent an assertion of an event of interest (fire) at a given location and time. Ground truth is obtained when the authorities choose to respond to or investigate a report. For example, if a citizen reports a hidden weapons cache, upon investigation, the cache will prove either to be present or absent.

To use the reputation system, we assume the citizens are asked to attach certainty to their reports. In practice, this information could be collected on a qualitative scale (e.g. from "very unsure" to "very sure"), and mapped onto quantitative values. In the simulation, the certainty is based on the citizen's distance from the event when they first observe it (the further away, the less certain). Unfriendly citizens (foes) can create false reports, and they may collude together to create a calling pattern that resembles a true report. The performance of the CER system is measured by the number of events responded to (fires extinguished) less non-events responded to (false alarms), divided by the total number of fires. The maximum performance is 1.

With anonymous reporting, the decision maker (DM) cannot discriminate true and false reports on the basis of reporter identity, greatly limiting the decision rules that can be implemented. There are various mechanisms for assuring calling identity using mobile devices, which will not be discussed here. For our purposes, we assume that personal reputations can be learned and factored into decision making.

The reputation system itself is implemented using the quadratic scoring rules (Eq. 4). The decision maker does not know which reports are true or false, but discovers ground truth only when the DM chooses to respond. At that point, each caller associated with the event receives reputation points. The decision to respond is

taken when there is a report from any citizen with a positive reputation, with certainty exceeding the prior by 0.1. The prior is the probability of an active fire at any grid location at any time.

Unlike a black list, which was investigated earlier [5], the reputation system gives citizens the ability to "earn" their way out of a negative reputation by making accurate reports. Thus, an honest reporter who unintentionally makes an erroneous report suffers a temporary (rather than permanent) loss of reputation.

The simulation was run with 30% foes, with 50% of reports from foes involving collusion. With reputation system, the number of false alarms dropped by 95%, and the performance (defined above) increased from 0.73 to 0.95. This shows that a reputation system can greatly enhance the performance of a CER system, even in the presence of a large contingent of foes determined to undermine the system.

5. Summary and Open Problems

We have presented and illustrated a reputation scoring approach that considers prior probabilities, so that scores combine accuracy and novelty to the consumer. Extending prior work on proper scoring functions [2, 3], the approach encourages expectation-maximizing providers to give accurate probabilities, while addressing the need to keep administrative effort small. We discussed ways to normalize the scores, in order to judge providers' accuracy, novelty, or productivity. The approach can succeed in situations where one can ascertain ground truth for a significant number of a provider's assertions.

We demonstrated the reputation system in the context of citizen event reporting, where information is collected from many potentially unreliable sources. A reputation system is clearly needed to help decision makers identify reliable and unreliable reporters. Soliciting a degree of certainty with each report encourages citizens to provide information even if they are not 100% sure of the facts. On one hand, they have the opportunity to express when they believe something to be certain. On the other hand, they can safely transmit uncertain knowledge by declaring a low level of certainty, reducing the fear of being punished for misleading authorities, if the tip turns out to be false. Hence more tips will be gathered, allowing the decision maker additional chances to create actionable information. Over time, those that express appropriate levels of certainty will become the most reputable information sources.

For future work, we believe that normalization may be important for some purposes, to adjust for different workloads – easier and harder questions, priors' being accurate (so agreement and zero novelty are optimal) versus inaccurate, and priors near 0.5 (no information) or .9 and .1 (unlikely to make large changes). It would also be desirable to explore scoring based on a logarithmic scale (so it is significant to move from .99 to .999 probability).

Aside from symmetry under negation, we have not considered the coherence of our scoring scheme under Boolean connectives (and, or). For example, an analyst could predict A, B, "A and B", and "A or B" at the same time, and there should be some relationship between the scores received for the related assertions.

How reputation translates into personal trust, and how those trusts are converted to beliefs and actions, is also open to investigation. Use of trust scores to synthesize multiple sources of information has been investigated in [7] and by other authors, but we believe a stronger and more direct link to reputation is needed.

References

1. Bayarri, M. J., Berger, J.: The Interplay Between Bayesian and Frequentist Analysis. Statistical Science, 19: 58-80 (2004)
2. Brier, G. W.: Verification of forecasts expressed in terms of probability. Monthly Weather Review, 75, 1-3 (1950)
3. Gneiting, T., Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation. J. American Statistical Assn. 102, 359-378 (2007)
4. Hogarth, R.: Cognitive Processes and the Assessment of Subjective Probability Distributions. J. American Statistical Assn. 70, 271 -289 (1975)
5. Kramer, M., Costello, R., Griffith, J.: Investigating the Force Multiplier Effect of Citizen Event Reporting by Social Simulation. Conference of the European Social Simulation Association (2008)
6. Savage, L: Elicitation of Personal Probabilities and Expectations. J. American Statistical Assn. 66, 783-801 (1971)
7. Zuo, Y., Panda, B.: Information trustworthiness evaluation based on trust combination ACM symposium on Applied computing, 1880-1885 (2006)

Aggregating Reputation Feedback

Florent Garcin[1], Boi Faltings[1], and Radu Jurca[2]

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL),
Artificial Intelligence Laboratory
{florent.garcin,boi.faltings}@epfl.ch
[2] Google Inc.
Switzerland
radu.jurca@gmail.com

Abstract. A fundamental task in reputation systems is to aggregate multiple feedback ratings into a single value that can be used to compare the reputation of different entities. Feedback is most commonly aggregated using the arithmetic mean. However, the mean is quite susceptible to outliers and biases, and thus may not be the most informative aggregate of the reports. We consider three criteria to assess the quality of an aggregator: the informativeness, the robustness and the strategyproofness, and analyze how different aggregators, in particular the mean, weighted mean, median and mode, perform with respect to these criteria. The results show that the arithmetic mean may not always be the best choice.

1 Introduction

Many sites on the world wide web offer people the possibility to share their experiences with products and services through reviews and ratings. This feedback helps people avoid bad choices, drives them towards more useful products, and brings more revenue to good producers. They are an important part of the user's decision making when buying goods or services.

We consider in particular reputation and rating systems for products and services, such as those operated by Amazon.com, Tripadvisor, and many other electronic commerce sites. These have the following characteristics:

- they collect ratings for individual well-identified products or services, and aggregate these ratings into a single score;
- the identity of raters does not have to be known;
- raters act to influence the score of the item they rate to make it as close as possible to the value they consider best. Note that this value might not reflect the true quality if the rater is not honest.

A common reflex for users of such sites is to order the choices according to their ratings, and only consider those that are at the top of such rankings. Such an order is usually obtained by aggregating individual feedback scores into a single value that establishes an ordering of the alternatives. The most common

way of aggregating ratings is by the arithmetic mean, but one can also consider aggregation using the weighted mean, the median and the mode.

In unbiased normal distributions, there is little difference between these aggregators. However, it is known that in reality, reviews are often biased [7]. Writing a review or even just leaving a rating requires effort, and since it is voluntary many of these ratings are left by people who have some ulterior motive or extreme opinion. One can thus observe that the distribution of ratings is far from the normal distribution one would expect from an unbiased population of raters. This means that the different ways of aggregating them can give very different results.

In this paper, we consider how the aggregation method influences the quality of the ranking. We evaluate quality using the following three criteria:

- informativeness, i.e. how likely is it that the ranking that a user finds at the time of making a choice will still be the ranking when the user is using the product or service;
- robustness, i.e. how easy is it for the ranking to be distorted by outliers or malicious reviews;
- strategyproofness, i.e. for a rater who wants the average ranking to be a certain value, is it best to report this value or manipulate the aggregation by reporting differently.

We compare four different ways of aggregating n numerical ratings $r_1, ..., r_n$, using different forms of averaging:

- the *mean* is the arithmetic mean $\overline{r_a} = \frac{1}{n}\sum_{i=1}^{n} r_i$.
- the *weighted mean* is the same as the arithmetic mean but with weights: $\overline{r_w} = \frac{\sum_{i=1}^{n} w(i)r_i}{\sum_{i=1}^{n} w(i)}$, where $w(i)$ is the weight function.
- the *median* is the smallest value $\overline{r_d}$ such that half of the values are $\geq \overline{r_d}$ and half of the values are $\leq \overline{r_d}$, i.e. there exist $X \subset \{r_1, .., r_n\}$ and $Y \subset \{r_1, .., r_n\}$ such that $(\forall r_i \in X) r_i \leq \overline{r_d}$ and $(\forall r_i \in Y) r_i \geq \overline{r_d}$ and $||X| - |Y|| \leq 1$.
- the *mode* is the smallest value $\overline{r_o}$ which occurs most frequently as a rating, i.e. for any $r' \neq \overline{r_o}$, $|\{r_i | r_i = \overline{r_o}\}| \geq |\{r_i | r_i = r'\}|$.

Both the median and the mode require a tie-breaking rule. When two values are possible, we select the smallest one. Moreover, it happens that two items have the same aggregated value. In that case, we use the number of reviews as a tie-breaking rule to make the final ranking.

We examine the reviews given on an actual review site and observe that the four notions of average differ significantly. In particular, the mode and median tend to be more robust to outliers and biased reviews than the mean and the weighted mean, and thus may be more informative for a user.

In this paper, we first present an analysis of the four different notions with respect to their robustness, and show that they have very different properties. We then analyze their behavior on data taken from an actual review web site, and show that they lead to very different rankings and also very different behavior of the rankings over time. In particular, our results suggest that the mean may

not be the most informative way of aggregating ratings since the ranking shown to a user is often very unstable.

2 Related Work

There are many reputation mechanisms that follow the model we assume in this paper. They can differ significantly in the way they aggregate and display reputation information to the users. Some mechanisms accumulate all reports into a reputation score that may potentially grow forever. eBay[3] and RentACoder[4] are two commercial sites where part of the reputation information is given by scores that reflect the total number of positive or negative interactions reported for an agent.

Amazon[5] or the popular movie review database IMDB[6] rank products by the arithmetic mean of ratings. They also publish histograms of the ratings[7] but the richer information is more difficult to find through the normal user interface, and is not used in any way for ranking the alternatives.

Tripadvisor[8] ranks hotels from cities around the world. The hotels are sorted by "popularity", defined here as the arithmetic mean of ratings. The reviews for a given hotel are ordered from the most recent to the oldest. Only 10 reviews are listed per page and contain information about the author (date, username, location of the reviewer) and the overall rating with a textual comment. The user has to click on the review to see more details.

Other mechanisms use discrete values for reputation information and define clear rules describing how sets of feedback are mapped to reputation values. The popular IT news site Slashdot[9] uses *karma* levels (i.e., *terrible, bad, neutral, positive, good, and excellent*) that characterize the quality of the news submissions posted by a user so far. Likewise, eBay sellers also have labels (e.g., *power seller*) that they can gain by meeting certain conditions.

The robustness of the reputation mechanism has also been an important concern of the research community. [5] discuss the risks associated with cheap *online pseudonyms* (i.e., users can easily create several online identities) and conclude that in any reputation mechanism newcomers must start with the lowest possible reputation. This property is later used by [3] to design moral hazard reputation mechanisms that are robust to identity changes.

[2] describes general techniques for making online feedback mechanisms immune to manipulation. A theoretical study of opinion manipulation is presented in [4], with the striking conclusion that manipulation can both increase and

[3] www.ebay.com
[4] www.rentacoder.com
[5] www.amazon.com
[6] www.imdb.com
[7] In addition, IMDB correlates demographics information with the histogram of scores.
[8] www.tripadvisor.com
[9] www.slashdot.org

decrease the information value of online forums. Other works addressing the robustness of the reputation information are [12] and [1].

[10] and [8] discuss general mechanisms for making reputation mechanisms incentive compatible. The idea is to reward the agents for reporting feedback such that the expected reward is maximized by being honest. [9] extends this idea to mechanisms that are also collusion resistant.

Our work differs from the above results in several important ways. First, we are looking at typical review forums where the social network of a user is unknown, and most users submit only one review. Second, we are looking at *single value* aggregators of reputation information, that can be easily understood and used by normal users to rank alternatives. Finally, we consider actual reviews and study how different information aggregators affect key properties like robustness, informativeness and strategyproofness.

3 Empirical Study

We consider feedbacks from a popular travel site that collects reviews of hotels from users around the world. The reviews contain a textual comment with a title, an overall rating and numerical ratings from 1 (lowest) to 5 (highest) for different features such as cleanliness, service, location, etc. The site provides ranking of hotels according to their location. Like most of the reputation sites, it aggregates reviews into a single value for each hotel and, based on that value, sorts hotels in ascending order. It uses a simple arithmetic mean on the overall ratings to recommend hotels.

We selected four cities for this study: Boston, Las Vegas, New York and Sydney. For each city, we took the first 100 hotels that have the highest number of reviews. Table 1 shows for each city the number of reviews and the distribution of hotels with respect to the star-rating provided by the website. Hotels that do not have a star-rating are classified as 'NA'. All data were collected by crawling the website in July 2007.

Table 1. A summary of the data set.

City	# Reviews	# of Hotels with NA, 1, 2, 3, 4 & 5 stars
Boston	5537	17, 2, 4, 23, 15, 5
Las Vegas	28017	19, 8, 18, 31, 17, 7
New York	29123	16, 9, 12, 35, 24, 4
Sydney	3629	41, 0, 1, 29, 19, 10

4 Robustness

In this section, we present an analysis of four aggregators, namely the mean, the weighted mean, the median and the mode, inspired by the robust statistics

theory. Robust statistics aim at analyzing and suggesting estimators that are unaffected by small deviations from the model assumptions. Interested readers may refer to [6] [14] for additional informations.

For this analysis, we quantify how robust aggregators are against outliers and malicious reports. In order to assess the quality of each aggregator, we define the *breakdown point* as a measure of this robustness. The breakdown point is the proportion of manipulated ratings required to make the aggregator return an arbitrary value.

Definition 1. *Let* $\{r_1, ..., r_{n-l}, r'_1, ..., r'_l\}$ *be a sample of* n *reviews where* r'_i *are outliers. The* finite-sample breakdown point ϵ *of an aggregator* $\overline{r}$ *is the smallest proportion* $\frac{l}{n}$ *for which the set* $\{r'_1, ..., r'_l\}$ *will cause* $\overline{r}$ *to be unbounded.*

Definition 2. *The* breakdown point ϵ^* *is the limit of the finite-sample breakdown point as* n *goes to infinity.*

This definition provides a tool to measure the robustness of every estimator. The higher the breakdown point, the more robust the estimator is. However, the breakdown point cannot exceed 0.5 because if more than half of the ratings are outliers, it is not possible anymore to distinguish the underlying distribution of the outliers. We will see in the next sections that two aggregators achieve this upper bound.

4.1 Mean and Weighted Mean

Let $\{r_1, ..., r_{n-l}, r'_1, ..., r'_l\}$ be the set of ratings for a given hotel h. r'_i are the outliers. We define the mean by

$$\overline{r_a} = \frac{1}{n}(\sum_{i=1}^{n-l} r_i + \sum_{i=1}^{l} r'_i) \tag{1}$$

One outlier is enough to change the value of the mean. Thus, the finite-sample breakdown point of the mean is $\epsilon = \frac{1}{n}$. The breakdown point is $\epsilon^* = \lim_{n\to\infty} \frac{1}{n} = 0$. The mean is extremely sensitive to outliers.

In this study, the ratings are bounded and we need more than one outlier to significantly alter the mean and thus the ranking. For that reason, we would like to quantify how many outliers are required to change the ranking of hotel h_j from position j to position i. Consider a hotel h_i with n ratings of mean $\overline{r_i}$. We add k outliers with mean $\overline{r}'$. How many outliers are needed to have a new mean lower or equal to the mean $\overline{r_j}$ of another hotel h_j? That is

$$\frac{n\overline{r_i} + k\overline{r}'}{n+k} \leq \overline{r_j} \tag{2}$$

After reordering, we get

$$k \geq \frac{n(\overline{r_i} - \overline{r_j})}{\overline{r_j} - \overline{r}'} \tag{3}$$

For instance, if we want to set the mean to 4 of a hotel with $n = 100$ ratings and mean $\overline{r} = 4.5$ by only adding lowest ratings of '1', then we need $k \geq 17$ outliers.

The weighted mean is similar to the simple mean except that ratings have assigned weights and some contribute more than others. Let $\{r'_1, ..., r'_l, r_1, ..., r_{n-l}\}$ be the set of ratings sorted from the most recent to the oldest for a given hotel h. The weighted mean is

$$\overline{r_w} = \frac{\sum_{i=1}^{l} w(i) r'_i + \sum_{i=1}^{n-l} w(i+l) r_i}{\sum_{i=1}^{n} w(i)} \tag{4}$$

The weights do not change the breakdown point and remains the same as the simple mean. Note that the mean is a special case of the weighted mean where the weights are all equal to 1. Obviously, the number of outliers needed to change the ranking of a hotel is upper bounded by Equation 3 and depends on the weight function $w(i)$.

4.2 Median

The median is the rating $\overline{r_d}$ separating the lower half from the upper half of a set of ratings. Let $\{r_1, ..., r_{n-m}, r'_1, ..., r'_m\}$ be the set of ratings sorted in ascending order for a given hotel h. r'_i are the outliers. If n is odd, that is $n = 2l + 1$, the median is located at $(l + 1)/n$. Recall that if n is even, i.e. $n = 2l$, we take the value at l.

To find the breakdown point, we determine the proportion of outliers required to change the value of the median. The finite-sample breakdown point is given by

$$\epsilon = \begin{cases} \frac{l+1}{n} = \frac{1}{2} + \frac{1}{2n}, & n = 2l + 1 \\ \frac{l}{n} = \frac{1}{2}, & n = 2l \end{cases} \tag{5}$$

Therefore, the breakdown point is $\epsilon^* = \lim_{n\to\infty} \epsilon = \frac{1}{2}$. The median is thus a robust aggregation function because it involves only the location and not the value of the ratings. To find the number of outliers required to change the ranking of a given hotel h with n ratings, we add k outliers to the ratings of h. In the worst case scenario, k should be at least equal to $n+1$. The first outlier determines the value of the median and thus the rank. For instance, if we want to change the median of a hotel that has 100 ratings, we need at least 101 malicious ratings and the new median is given by the first malicious ratings we introduce.

4.3 Mode

The mode, denoted $\overline{r_o}$, is another aggregation function and is equal to the rating that occurs the most frequently. That is, for any $r' \neq \overline{r_o}$,

$$|\{r_i | r_i = \overline{r_o}\}| \geq |\{r_i | r_i = r'\}| \tag{6}$$

Let m and l be the number of identical ratings r_1 and r_2 respectively, $m \neq l$. Obviously, if the mode is the rating r_1, then $m > l$. Therefore, $m \geq l + 1$. Thus, the finite-sample breakdown point of the mode is equal to $((n/2) + 1)/n$ for $n = m + l$ ratings. It follows that the breakdown point $\epsilon^* = \frac{1}{2}$. From the same reasoning, we need $k = n + 1$ outliers of the same value to change the mode. For instance, if a hotel as a mode $\overline{r_o} = 4$ with $n = 100$ ratings (of '4'), $k = 101$ outliers are required to change that mode.

5 Empirical Results

It is well-known that distributions of reports are far from normal due to reporting biases [7]. Aggregators such as the mean, median and mode have relatively the same value for normal distributions. However, they should have a significant difference for non-normal distributions. To support this hypothesis, we conducted the following experiment. For each of the four cities considered in our study, we computed a full ranking of the hotels according to each of the four aggregators explained in Section 1. Then, for every pair of aggregators we measured the *distance* between the corresponding orderings of hotels within a city. To measure the distance between the two rankings we chose the average absolute difference between the position of the same hotel in the two rankings.

For the weighted mean, we use Equation 7 as the weight function that is directly inspired by the logistic function applied in regression models. With this function (see Fig. 1), recent ratings have a high weight and the weight decreases while the rating is getting older. We use the following logit model for the relevance and thus the weight of a rating as a function of its order:

$$w(i) = \frac{0.9}{1 + e^{\beta(i-\mu)}} + 0.1 \tag{7}$$

Such logit models are commonly believed to be good models for probabilities that vary over time or space.

The results are presented in Table 2. For example, the rank of a hotel in Boston varies on the average with 7.7 positions (up or down) when the ranking is done according to the median instead of the mean. Likewise, the rank of a hotel in New York varies with an average of 16.9 positions (up or down) when the ranking considers the mode instead of the mean.

The average difference of ranks triggered by different aggregators is quite high: 8 to 17 ranks[10]. Considering that most feedback websites display only the first 5 or 10 "best" items, the results of Table 2 show that different aggregators can completely change the list of candidates suggested to the users. It therefore becomes important to better understand the properties of each aggregator.

[10] The only exception is the tuple *mean - weighted mean*. The two aggregators are conceptually very close, therefore the rankings span by them are also similar.

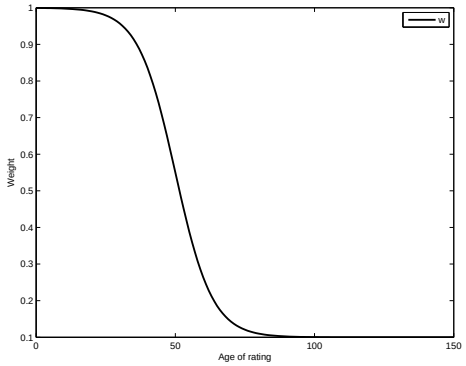


Fig. 1. The weight function: recent ratings (low index) get a highest weight. Note that the age is not related to the time but to the most recent rating.

Table 2. Average difference of ranking for the three aggregator functions.

	Boston	Las Vegas	New York	Sydney	*average*
mean - median	7.788	13.480	11.480	9.100	10.462
mean - mode	9.939	15.100	16.980	11.420	13.360
mean - weighted mean	2.394	2.760	5.480	1.340	2.993
median - mode	10.182	16.140	16.860	10.340	13.380
median - weighted mean	8.333	13.460	12.600	9.600	10.998
mode - weighted mean	10.848	15.740	17.940	11.700	14.057

5.1 Informativeness

In a reputation system, the goal of the aggregator is to reflect the user's reviews into one value. One assumption of aggregator is that users have reported their true experience. However, it is often not the case. For instance, the ratings are often part of discussion threads where past reviews influence future reports by creating prior expectations [13]. Therefore we can ask how an aggregator will continue to correctly reflect users' opinion. In Table 3, we look at the stability of each aggregator by counting the number of rankings that deviate by more than two ranks from the final ranking. The median is the most stable aggregator with two cities. However, the weighted mean seems more stable on average. The median follows closely. Then the mode and the mean come after.

As an example, Figure 2 provides the evolution of ranking and rating by the incoming reviews for a New York hotel. If we look at the mean aggregator only, when the rating decreases, the hotel loses ranks. However, around the 120th review, the rating increases and thus the hotel is going up in the global ranking. Although the median and the mode have a fixed value for the rating, the rank oscillates a little bit for the first reviews to stabilize very quickly. We observe such behavior for most of the hotels in our database.

Table 3. Average number of ranking that deviate from the final ranking with more than 2 ranks. In bold, the lowest value. The significance levels are computed with a one-way analysis of variance.

	Boston	Las Vegas	New York	Sydney
Weighted mean	29.606	**154.640**	96.660	23.450
Mean	45.833	227.120	156.930	28.460
Median	**23.652**	189.770	91.870	**12.760**
Mode	29.758	254.170	**73.550**	17.330
p-value	0.000	0.006	0.000	0.000

5.2 Robustness

Finally, we look at the robustness of each aggregator by taking the number of outliers required to alter the ranking of a given hotel. For each hotel, we inject outliers with the highest possible ratings, i.e. 5, until the rank changes. Table 4 summarizes the results for each city. Two reviews are enough to change the rank when the aggregator is the weighted mean, around 5 for the mean while the median and mode need 20 and 15 outliers respectively. The mean and weighted mean can be changed with a very low number of additional ratings. However, the mode, and in particular the median require a relatively large number of additional ratings, and are thus more difficult to manipulate.

Table 4. Average number of outliers (with highest ratings '5') required to alter the ranking. In bold, the highest value. The significance levels are computed with a one-way analysis of variance.

	Boston	Las Vegas	New York	Sydney
Weighted mean	1.922	2.153	2.155	1.464
Mean	3.328	5.102	8.041	1.948
Median	**10.297**	**40.602**	**22.639**	3.639
Mode	9.047	23.867	22.309	**3.691**
p-value	0.000	0.000	0.000	0.000

6 Strategyproofness

Besides influencing the conclusions that are drawn from a given set of ratings, the way that ratings are aggregated can also have an influence on the reports that users will submit. In this section, we consider to what degree users have an incentive to report a rating that differs from their true perception in order to manipulate the ranking.

We make the assumption that a rater has a single most preferred score that she would like to see as the aggregated score of the item being rated. For an honest rater, this value should be the true perception of quality. We furthermore

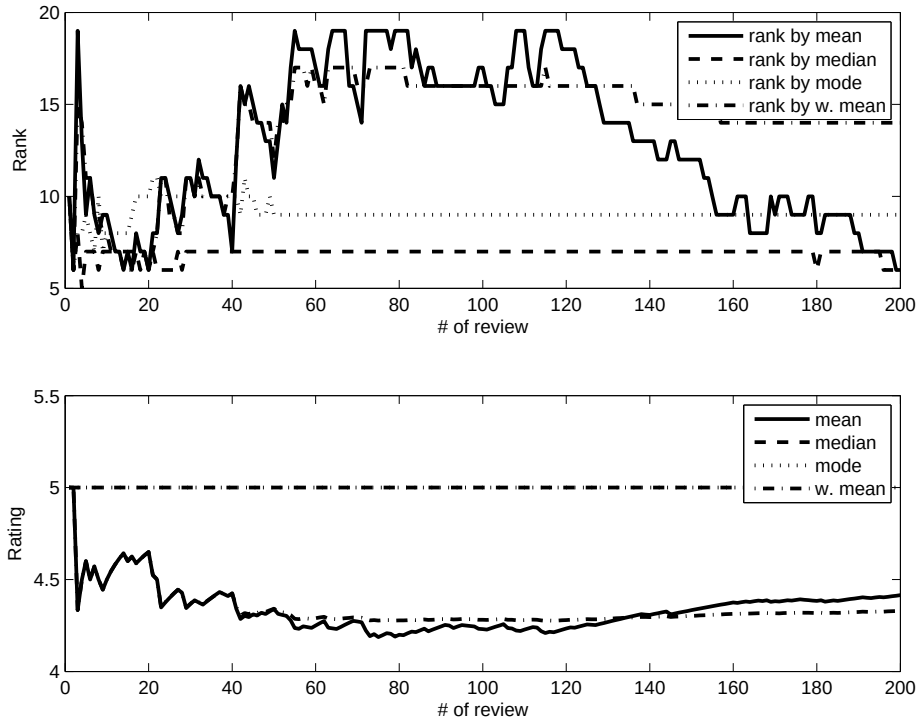


Fig. 2. A New York hotel

assume that when it is not possible to make the aggregated score take this most preferred score, the rater would like to bring it as close as possible to it. In the language of decision theory, this means that raters have a *single-peaked* preference profile: their preference for different ratings has a single peak at their most preferred score and drops monotonically to both sides of it.

Now consider the rating that such a user should report to best achieve its objective. In a *strategyproof* reputation system, a rater can expect the best possible outcome by reporting her most preferred score.[11] However, this is not always the case. For example, if a product currently has 5 reports with an arithmetic mean of 4, and the rater would like to see a score of 3, then it would be best off to report 1 and drive the mean to 3.5 rather than 3 and obtain a mean of 3.833. We believe that such manipulation strategies are the source of much of the reporting bias we can observe in practical reputation sites, and we conjecture that much more useful information could be obtained if the systems were indeed strategyproof.

Definition 3. *An aggregation function is* strategyproof *(or truthful) if there is no incentive for any of the reviewers to lie about or hide their private valuation.*

[11] Note that this does not have to be the true quality.

6.1 Mean and Weighted Mean

The mean and the weighted mean are not strategyproof. Consider that the reviewers are sorted in order of their private opinions. Let $\bar{r}$ be the mean (or weighted mean). Any reviewer a_j with a private opinion below the mean has the incentive to submit an exaggerated negative review in order to push the mean downwards. Likewise, a rater with a private opinion higher than the mean has the incentive to submit an exaggerated positive review.

6.2 Median

Moulin proves that, when preferences are single-peaked along the real line, the median is the only strategyproof preference aggregation scheme [11]. Assume that the reviewers $\{a_1, ..., a_n\}$ are sorted increasingly according to their private opinion of a hotel A. Let r_i be the private opinion of the reviewer a_i, so $r_{i+1} \geq r_i$. Let r^* denote the median rating, corresponding to reviewer a_{i^*}. Obviously, reviewer a_{i^*} should not deviate. If a reviewer a_j with $j < i^*$ misreports a lower value than r_j the median rating will not change. Misreporting a value higher than r_j, on the other hand, can only increase the median, and therefore make the public reputation of the hotel even further from a_j's private opinion. The same argument applies for any reviewer a_j with $j > i^*$. As long as the tie-breaking is independent of the reviews, then the same argument holds even if there is an even number of raters in the system.

In addition, Moulin ([11]) also shows that aggregation through the median is Pareto optimal and anonymous.

6.3 Mode

The mode is not strategyproof. Assume that two reviewers have the same private opinion r_1 and three reviewers have same private opinion $r_2 > r_1$. Let a_j be a reviewer whose private opinion is $r_j < r_1$. If the reviewer misreports and submits a review with the value r_1 she has successfully modified the public reputation of the hotel from r_2 to r_1, which is a better outcome for a_j.

7 Conclusion

We considered different ways of aggregating ratings, in particular the mean, weighted mean, median and mode. All review sites that we are aware of aggregate ratings by taking their mean, and if ratings were unbiased and normally distributed the different notions would not differ much. However, in actual review sites there are many biases, and the three methods give very different results. In hindsight, we find it surprising that other ways of forming averages have not been considered.

We considered three criteria: informativeness as reflected by the degree to which the ranking fluctuates over time, robustness as reflected by the number

of reports necessary to change the aggregate, and strategyproofness as reflected by the incentive to file truthful reports to move the average as close to them as possible.

On all three criteria, the mean seems to be the worst way of aggregating rankings: it changes the most frequently, it is the least robust, and it is not strategyproof. While the weighted mean is in general more informative, the median is significantly more robust. Finally, only the median is strategyproof. Strategyproofness may greatly increase the quality of rating information that is collected, provided that raters actually understand it. This would be an interesting subject for a user study.

We thus conclude that for using reputation sites to help users in their choices, aggregation through the median or mode are likely to be better choices than the mean. However, we recognize that if the purpose of the reputation system is to encourage good quality, i.e. to deal with the moral hazard problem, it may actually be desirable for raters to be able to move the ranking easily. The two aspects should be weighed by the designer of a reputation system.

Acknowledgments

The authors would like to thank David Parkes for pointing us to [11] and the anonymous reviewers for their contructive comments.

References

1. S. Buchegger and J.-Y. Le Boudec. The Effect of Rumour Spreading in Reputation Systems for Mobile Ad-hoc Networks. In *WiOpt '03: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, Sophia-Antipolis, France, 2003.
2. C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *EC '00: Proceedings of the 2nd ACM conference on Electronic commerce*, pages 150–157, New York, NY, USA, 2000. ACM.
3. C. Dellarocas. Reputation Mechanism Design in Online Trading Environments with Pure Moral Hazard. *Information Systems Research*, 16(2):209–230, 2005.
4. C. Dellarocas. Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms. *Management Science*, 52(10):1577–1593, 2006.
5. E. Friedman and P. Resnick. The Social Cost of Cheap Pseudonyms. *Journal of Economics and Management Strategy*, 10(2):173–199, 2001.
6. F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, 1986.
7. N. Hu, P. Pavlou, and J. Zhang. Can Online Reviews Reveal a Product's True Quality? In *Proceedings of ACM Conference on Electronic Commerce (EC 06)*, 2006.
8. R. Jurca and B. Faltings. Minimum Payments that Reward Honest Reputation Feedback. In *Proceedings of the ACM Conference on Electronic Commerce (EC'06)*, pages 190–199, Ann Arbor, Michigan, USA, June 11–15 2006.
9. R. Jurca and B. Faltings. Collusion Resistant, Incentive Compatible Feedback Payments. In *Proceedings of the ACM Conference on Electronic Commerce (EC'07)*, pages 200–209, San Diego, USA, June 11–15 2007.

10. N. Miller, P. Resnick, and R. Zeckhauser. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science*, 51:1359 –1373, 2005.
11. H. Moulin. On strategy-proofness and single peakedness. *Public Choice*, 35:437–455, 1980.
12. M. Srivatsa, L. Xiong, and L. Liu. TrustGuard: Countering Vulnerabilities in Reputation Management for Decentralized Networks. In *Proceedings of the World Wide Web Conference*, Japan, 2005.
13. A. Talwar, R. Jurca, and B. Faltings. Understanding user behavior in online feedback reporting. In *EC '07: Proceedings of the 8th ACM conference on Electronic commerce*, pages 134–142, New York, NY, USA, 2007. ACM.
14. R. R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Elsevier Academic Press, 2nd edition edition, 2005.

3 Field and Laboratory Research on Reputation

Running head: WHERE DOES REPUTATIONAL POWER IN ORGANIZATIONS COME FROM

Where does Reputational Power in Organizations Come From?[1]

Alona Labun[2]
University of Groningen

Rafael Wittek, Christian Steglich, Rudi Wielers
University of Groningen

December 15, 2008
(5295 words)

[1] Paper prepared for the First International Conference on Reputation: Theory and Technology, March 18-20, 2009, Gargonza, Italy.

[2] Corresponding author. Address: Department of Sociology/Graduate School ICS, University of Groningen, Grote Rozenstraat 31, 9712 TG Groningen. Email: A.Labun@rug.nl.

Where does Reputational Power in Organizations Come From

Abstract

In this paper we develop and empirically test a model of the dynamics of perceived informal power in organizations. We argue that individuals tend to attribute power to actors who are perceived as powerful by many others (imitation). We further suggest that interpersonal trust relations with a specific individual increase the likelihood of power attributions to this person. Emphasizing instrumental motives and the importance of brokerage for power acquisition, we also argue that employees using indirect influence strategies are more likely to be perceived as powerful. However, emphasizing affective motives, we suggest that the use of direct strategies like negotiation and persuasion increases perceived power. We apply stochastic, actor-based models for network evolution to longitudinal social network data (4 waves) collected in the management team of a German Paper Factory (n=17). Results show strong effects of imitation and interpersonal trust, and support for the negative effect of indirect formal strategies.

Key words: perceived informal power, reputation, trust relation, influence strategies, network dynamics.

Where does Reputational Power in Organizations Come From

Introduction

The purpose of this article is to develop and empirically test a model of the perceived informal power differences in organizations. Power since long occupies a central position on the research agenda of organizational scholars, who also acknowledge the importance of distinguishing between formal and informal power on the one hand (Brass, 1984; Krackhardt, 1990), and between objective and perceived power differences on the other hand (Bacharach & Lawler, 1976; Gioa & Sims, 1983; Fiol, O'Connor & Aguinis, 2001).

Informal power has been variously defined as the ability to influence others and get things done (Brass, 1984; Emerson, 1962), to mobilize resources (Roberts, 1986), or to lead through "personal appeal" (Krackhardt, 1990). Power perceptions (Meindl, Ehrlich, & Dukerich, 1985; Pastor, Meindl, & Mayo, 2002; Pfeffer, 1977) or power mental models (Fiol et al., 2001:225) refer to "organized mental representations of one's own and others' power". More specifically, reputational or perceived power has been defined as the set of beliefs others hold about how powerful an actor is (ibid). Hence, an actor is powerful when seen as powerful by others. Certain organizational members may appear to be influential in the eyes of others, whereas others may be perceived as rather powerless.

The study of the perceived informal power differences is important for several reasons. First, variations in power can have a tremendous impact on the success or failure of many organizational processes, including the individual level, group level, and organizational level outcomes. On one hand, deviations from formal lines of command undermine the legitimate lines of communication and authority, and thus may be detrimental for the functioning of the organization. Disproportionate influence of one individual can also have disruptive consequences on the team process by reintroducing hierarchy. On the other hand, the emergence of capable informal leaders can also be beneficial and may help resolve problems resulting from imperfections in the design of the formal structure (Cross & Prusak, 2002; Krackhardt & Hanson, 1993). Influential individuals are likely to control the resource flows and potential opportunities in organizations. Employees can reap rewards in terms of access to and control over diverse communications, distribution of ideas and other valued resources throughout their immediate social circle by knowing how much and what kinds of power others have. Second, employees usually do not respond to objective power, but to subjectively perceived power of other actors (Gioa & Sims, 1983). Hence, the

formally powerful are not always those who really have the power. Often individuals in low-ranking positions may be perceived as informally powerful and therefore be able to exert much more influence than formally powerful actors. This power is not associated with their formally defined positions within the organization (Aghion & Tirole, 1997; Ibarra, 1993). Third, in complex organizations, even formal power concerning specific issues or domains is not always clearly defined. Reputation may be most important when objective information on an actor's behavior is scarce or absent (Sharman, 2007). Employees will make use of behavioral, structural, and situational cues in order to infer who is powerful.

Despite the attention that has been paid to informal power and power perceptions separately, efforts to model the antecedents and dynamics of perceived differences in informal power reputation still are surprisingly scarce. Research on power perceptions so far has focused on the determinants of subordinates' perceptions of managerial power (Bacharach & Lawler, 1976; Gioa & Sims, 1983), leaving unaddressed how peers evaluate informal power differences among subordinates as well as between subordinates and superiors. Furthermore, most research on informal power has been cross-sectional in nature, with the result that little is known about how reputational power changes through time.

Hence, our study addresses the following two questions: Why are some organization members perceived to have more informal power than others? Which factors affect the stability or change of perceptions concerning someone's informal power position? Building on previous research in the field of social networks and power tactics, we argue that (changes in) perceived informal power of a focal actor is driven by other actor's perceptions of the focal actor's informal power (i.e., his power reputation), the focal actor's use of power strategies, and the focal actor's social embeddedness.

In what follows, we will first elaborate the theoretical background and derive testable hypotheses. Section three presents the research design and data. Section four presents the results. Section five concludes.

Theoretical Background

A crucial question for actors in organizations is how to assess who has power in their social environment. Organizational settings are characterized by a great deal of ambiguity. There is always a certain level of uncertainty concerning whether or not an

employee is actually powerful. As a consequence, in order to evaluate the relative informal power of a focal actor, individuals will have to make use of both direct and indirect cues as indicators of the focal actor's position. Direct cues can be drawn from own observations of a focal actor's behavior towards oneself or others. For example, by observing how a colleague successfully influences others in my team may affect my perception of this colleague's informal power. Indirect cues can result from information provided by others concerning a focal actor's informal power. For example, my assessment of a colleague's informal power may increase if my team mates regularly refer to her as the one who succeeded in getting her ideas implemented. More specifically, previous research identified three core mechanisms accounting for the emergence of perceived informal power differences in uncertain environments: rational imitation, power strategy use, and social ties. We will address each of them in turn.

Rational Imitation

Informal power differences and status hierarchies emerge in bilateral interactions between group members (Gould, 2002). Even in relatively small groups, there are limits to the horizons of observability (Friedkin, 1983): group members can not directly observe all interactions. People often work in different groups, and thus participate in different social worlds. Organizational members may be unaware of the social relations between other employees, and the extent to which a certain actor is influential in such relationships. Since one is never embedded in every possible interaction that is taking place in the organization, the available information regarding degree of influence of another actor is to a certain extent limited. In addition, power differences between exchange partners can change through time. Uncertainty about the relative distribution of informal power will be the result: individuals will have only partial information about a focal actor's relative position towards other members in the group. Individual judgments in this ambiguous situation are generally more open to the influence of others, hence one way to cope with the uncertainty is to rely on the power attributions of other actors in the system. Theoretical approaches ranging from rational choice (Hedstrom, 1998) and game theory (Barrera & Buskens, 2007) to Neo-Institutional organization theory (DiMaggio & Powell, 1983; Wittek, 2003) have pointed towards imitation as a viable strategy for dealing with uncertainty. Most of this research refers to assessment of trustworthiness of exchange partners. We

propose that imitation will also be an important strategy driving perceptions of informal power differences in groups.

In judging an actor's reputation individuals are likely to look for cues and associations from others in the group (Sharman, 2007). They scan the environment for potentially valuable information on (changes in) the relative power position of other group members. This information can reach them directly through personal observation or indirectly via the grapevine, i.e. through communicating with others.

Hypothesis 1 (Rational Imitation). The higher the number of group members who perceive a focal actor to be powerful within the group at t_0, the more likely it is that ego perceives the focal actor as powerful at t_1.

Interpersonal Ties

A second source of information about an actor's relative position in an informal power structure is personal ties to others. We emphasize personal ties since they are usually characterized by mutual trust and respect. Organizational members who have a personal tie with each other are more likely to trust each other, and hence share valuable information and advice. Furthermore, a personal tie to a focal actor can be an especially useful source of more complete information concerning the focal actors' social relations to other group members, as well as degree of influence he or she has in such relationships. The more I trust, the more likely I will get a positive impression of this person, and the more likely I will be disposed to believe this person's accounts relating to interactions with third parties (Hess & Hagen, 2006). Previous research has also suggested that individuals in organizations see themselves as more popular than they actually are (Kumbasar, Romney, & Batchelder, 1994). This bias is likely to play a role in the focal actor's communication with trusted alters. Persons linked to the focal actor by an interpersonal trust relationship have a higher likelihood to be exposed to the focal actor's (biased) accounts of social influence attempts. With interpersonal trust increasing the willingness to believe the trusted person's accounts, perceived power between persons linked by an interpersonal trust relation should be higher than perceived power of persons with whom no interpersonal trust relation exists.

Where does Reputational Power in Organizations Come From

Hypothesis 2 (Interpersonal ties). The stronger the interpersonal trust between an individual and a focal actor at t_0, the more likely it is that he or she will perceive the focal actor as powerful at t_1.

Power Strategies

A third potential indicator of an individual's informal power is behavioral cues. The use of power strategies has been shown to play a significant role in affecting power perceptions of others. Power strategies are a means by which an individual tries to accomplish his or her personal goals in a social environment, such as enhancing and maintaining his or her hierarchical position (Kyl-Heku & Buss, 1996; Lund, Tamnes, Moestue, Buss, & Vollrath, 2007). The use of influence strategies in itself may be regarded as a source of power, and hence is positively related to perceptions of power (e.g., Brass & Burkhardt, 1993; Gioa & Sims, 1983).

Building on the distinction between direct and indirect power strategies (Morrill & King Thomas, 1992), we elaborate two competing mechanisms. The first mechanism emphasizes affective motives and the importance of direct, bi-lateral strategies. Influence tactics based on negotiation and persuasion help to create and reinforce interpersonal bonds, as well as build and reproduce interpersonal trust. Furthermore, by using direct strategies an individual signals concern for the other and the relationship. According to this perspective, employees who are skillful in the use of direct influence strategies are more likely to elicit cooperative behavior from others, thereby building a reputation of someone who gets things done and can be trusted. The preference for direct strategies will be positively related to the actors' power reputation. The more one engages in direct, bi-lateral strategic behavior, the more powerful he will be perceived by others.

Hypothesis 3 (Direct power strategies). The stronger a focal actor's tendency to use direct power strategies at t_0, the more likely it is that ego perceives the focal actor as powerful at t_1.

The second perspective emphasizes instrumental motives and the importance of brokerage for power acquisition. Actors in brokerage positions need to engage in brokerage behavior to reap the benefits of their social network. Employees who are skillful in the use of indirect, tri-lateral strategies of influence (e.g., gossiping) are

able to manipulate information, and exploit information asymmetries. Furthermore, by using the strategies that are based on social display and networking, an actor signals ability to bring together his social contacts (e.g., mobilize a coalition) when this is likely to generate some advantage, thus enhancing a reputation of someone who is able to influence others and get informal support. The preference for indirect strategies will be positively related to the actors' power reputation. The more one engages in indirect, tri-lateral strategic behavior, the more powerful he will be perceived by others.

Hypothesis 4a (Indirect power strategies). The stronger a focal actor's tendency to use indirect power strategies at t_0, the more likely it is that ego perceives the focal actor as powerful at t_1.

The use of indirect strategies that involve high-ranking organization members as a third party signals association with "winners", with the successful and the powerful. Previous research has shown that individuals' personal reputations can be enhanced by the mere perception that one is socially connected to prominent others (e.g., Kilduff & Krackhardt, 1994). By basking in the superiors' reflected glory, one may build a reputation of an influential actor (Pfeffer & Fong, 2005). Hence, we assume that the preference for indirect strategies involving superiors will be positively related to the actors' power reputation. The more one gets things done through the superior, the more powerful he will be perceived by others (basking in reflected glory effect).

Hypothesis 4b (Formal power strategies). The stronger a focal actor's tendency to use formal power strategies at t_0, the more likely it is that ego perceives the focal actor as powerful at t_1.

On the other hand, indirect strategies involving superiors can also be detrimental for one's power reputation among other employees. The "basking-in-reflected-glory-effect" may also work in reverse in a sense that connections to formally powerful individuals who are viewed negatively, for example, could detract from one's reputation (e.g., Mehra, Dixon, Brass, & Robertson, 2006). Complaining to someone in a higher formal position, for example, may signal one's inability to get things done and may destroy interpersonal trust. Hence, the preference for indirect strategies

involving superiors will be negatively related to the actors' power reputation: the more one gets things done indirectly through the assistance of the higher-ranking organizational members, the less powerful he will be perceived by others.

Hypothesis 4c (Formal power strategies). The stronger a focal actor's tendency to use formal power strategies at t_0, the less likely it is that ego perceives the focal actor as powerful at t_1.

Data and Method

Investigating the mechanisms behind the evolution of perceived informal power in organizational settings requires sociometric choice data. It also requires a setting in which some substantial change in the formal and informal interaction patterns has taken place. Network panel data (four measurements with six months intervals between each wave) that were collected from the members of the management team of a German Paper Factory from late 1995 until mid-1997 (Wittek, 1999) meet these criteria.

The Organization

The organization was located in a village with 800 inhabitants in southern Germany. It had seven departments: production, the chemical lab, maintenance, logistics, personnel, technical customer service and a project department. There were 17 male managers with a mean age of 40.53 (range: 28-51; *SD* = 10.19) who had worked 11.59 years (*SD* = 11.8) for the organization. Two-thirds of the managers had a degree in engineering.

When fieldwork started in 1995, the factory had 170 employees and two paper machines. After a bankruptcy in 1993, the company was taken over by German multinational, which decided to invest 40 million German Marks to enlarge the site by adding a new production hall and a third paper machine. The latter was scheduled to be operative on 1 September 1995. This project and the realization of the deadline of 1 September 1995 comprised the most central event in the factory throughout the observation period. During this time the managers had to cope with a double workload. In addition to their individual function in the daily production process, they were now also responsible for the successful realization of the common project. Mutual interdependence between the managers and the necessity to coordinate and cooperate reached previously unknown heights. During this phase, a clear group goal

was present. With the successful completion of the project at the end of 1995, the common group goal disappeared, although the production department still formed a single entity. The allocation of responsibilities concerning the new paper machine was highly ambiguous. In the beginning of 1996, solving the new machine's implementation problems was, on the whole, considered to be a joint task. Finally, in 1997, the production department was split up into three semi-autonomous units.

The team in which most of the managers participated on a regular basis was characterized by stable membership, frequent unscheduled lateral and vertical communication, and weekly meetings. Evidence from participant observation and a survey confirms the self-perceptions of the team members as a highly solidary work unit operating on the basis of trust rather than hierarchical control (see Wittek, 1999, pp. 86-100, 122-134). All the managers described their team as a "trust culture".

Measures

Reputational Power. Individual power was assessed at four points in time by asking each respondent to indicate on a 5-point Likert scale ranging from 1 (very little influence) to 5 (very much influence) how much influence each colleague (from the presented name list) has in the work activities of the factory. The question was worded as follows: "In each team there are members who – due to their personality or experience – have more influence on collective decisions than others. Through their enthusiasm or charisma they succeed more often than others to convince their colleagues about their ideas, and to get their ideas implemented. In your opinion, how much influence does each of the colleagues in the following list have (including yourself)?" To model the dependent network observations with the module SIENA the reputational power measure had to be dichotomized. Scores ranging from 1 to 3 were coded as "0" and scores ranging from 4 to 5 were coded as "1".

Power Strategies. Three types of power strategies were measured: direct, indirect, and formal. Each of them was measured at two points in time (i.e., in wave one and wave three, respectively). The question referred to how respondents dealt with cooperation problems: "There are many ways how people deal with cooperation problems. How appropriate do you, personally, consider each of the following behaviors?". The question was followed by twelve statements, five of which are used in our analysis. Two items capture direct strategies: bilateral arguing ("To speak to the other person in private") and public negotiation ("To discuss the problem during a meeting"). Indirect

strategies were measured with two items: retaliation ("To pay the person back with his own medicine") and resignation ("To keep one's anger for oneself and do nothing"). Formal strategies were measured with one item: complaining to a superior ("To complain to the manager over colleague"). Respondents were asked to rate the appropriateness of each strategy on an interval scale ranging from -100 per cent "inappropriate" to +100 per cent "appropriate", respectively, on a bipolar scale (later recoded into scores between -1 and +1). Since the strategies were measured only twice, the values of wave two and wave four had to be imputed. Appropriateness of the strategies at the second time point was imputed using the reported values in the first time point. Similarly, appropriateness of the strategies at the fourth time point was imputed using the reported values in the third time point.

Interpersonal Trust. The interpersonal trust level was measured at four points in time by the following question: "We all feel closer to some colleagues than to others. By "closeness" we mean how strongly you trust a specific colleague. For example, who do you confide important personal information (private or work related) to? Please indicate on the following list of colleagues, which of the descriptions comes closest to your relationship with this colleague." The answer categories were: "Person not known to me", "Distant – you would not confide even unimportant personal matters to this person", "Neutral – you do not know this person well enough to confide personal matters to him", "Strong – you confide matters to this person that are relatively important to you" and "Very strong – you confide matters to this person that are very important to you".

Formal Position. Formal position was included as a control variable. Information on each manager's formal position was assessed based on organizational chart during fieldwork. It is coded as "1" if the respondent was a supervisor of at least one person in the network, and "0" otherwise.

Difference between Periods. Ethnographic evidence on the organization showed that during the second measurement, some major changes in the structure had been implemented, which seriously affected the site managers' power position, and also led to substantial changes in the informal relations. To control for the difference between periods a dummy period variable (coded "1" for the second period and "0" for the others) was included in the model.

Method

In order to study the dynamics of networks we applied recently developed stochastic actor-based network model (see Snijders, 2005) for the analysis of longitudinal data on social networks. The model is implemented in the SIENA – part of the StOCNET software package (Boer et al., 2006; Snijders, 2005; Snijders et al., 2007). The continuous model describes the development of a social network through time as a result of relational changes made over time by members of the network. The first network observation is itself not modeled but used only as a starting point of the simulations. The model estimates the behavior rules that fit best the observed trajectory of the networks. The network structure, individual attributes and dyadic covariates are taken into account. In addition to the parameters corresponding to the proposed hypotheses, a number of control parameters was included. As general control variables SIENA automatically includes the constant change rate (the amount of change between the two measurement moments) and density. Furthermore, we take into consideration difference between measurement periods and actors' formal position in the organization as control variables since both might have an impact on power attributions.

Results

Table 1 presents several descriptive statistics. The average degree of influence (i.e., the average number of people one perceives as influential) and the average degree of trust (i.e., the average number of people one trusts) show the substantial change in the informal interaction patterns that occurred through out the fieldwork period in the organization. The mean statistics of power strategies suggest that public negotiation is considered as the most appropriate means of dealing with cooperation problems, whereas retaliation appears to be the least appropriate strategic behavior.

- Table 1 about here -

The results of the analyses are summarized in Table 2. First, the results on the descriptive level will be addressed, and then those pertaining to our hypotheses. We will refer to the parameters by their numbers in the table, and first address the effects of the control variables.

- Table 2 about here -

Where does Reputational Power in Organizations Come From

The network rate parameters indicate that the highest estimated average amount of change per actor occurred from the third to the fourth time point at which the influence network was observed (period 3). The negative outdegree effect (4) indicates a low density of the network and has no substantial meaning otherwise. The significant dummy effect for period 2 (5) indicates that in the second period there was a particularly strong trend to withdraw informal power from actors in one's network. We also controlled for actors' formal position (8) in the organization. There was no significant effect of formal position on informal power over time, that is focal actors' formal power had no effect on whether or not over time he was perceived as informally powerful by others.

Turning to the hypotheses, the popularity effect (6) captures the idea that the more group members perceive a focal actor as powerful at t_0, the more likely it is that ego will see the focal actor as powerful at t_1. This parameter is significant and points into the predicted direction, thus lending support to our hypothesis concerning rational imitation (H1).

The interpersonal ties hypothesis (H2) posits that the more ego trusts the focal actor at t_0, the more likely it is that ego will perceive the focal actor as powerful at t_1. The trust effect (7) was significant and points in the predicted direction, supporting the interpersonal ties hypothesis.

The bilateral arguing and the public negotiation effects capture the idea that the stronger the actors' tendency to use direct power strategies at t_0, the more likely it is that this actor will be perceived as powerful by others at t_1. Both the bilateral arguing effect (9) and the public negotiation effect (10) are not significant: bilateral arguing and public negotiation do not play a role in power attribution process. These findings do not provide support for H3, according to which the use of direct power strategies increases a focal actors' perceived informal power.

According to the indirect power strategies hypothesis (H4a), individuals who use indirect strategies at t_0 are expected to be seen as powerful at t_1. The retaliation effect (11) is significant at the 10% level, but points into the opposite direction as predicted by the hypothesis. The resignation parameter (12) is not significant. This implies that H4a has to be rejected.

The two competing hypotheses concerning (indirect) formal power strategies suggest that one's tendency to use formal power strategies at t_0 may result in others perceiving the focal actor as more powerful (H4b) or as less powerful (H4c) at t_1. The

"complain to superior" parameter (13) was significant and has a negative sign. This finding lends support for H4c and disconfirms H4b: individuals activating formal authorities to influence others are less likely to be viewed as informally powerful by other group members.

Discussion and Conclusion

In the current study we addressed the following two questions: Why are some organization members perceived to have more informal power than others? Which factors affect the stability or change of perceptions concerning someone's informal power position?

Consistent with theories pertaining to rational imitation, we found that the higher the number of group members who perceive a focal actor to be powerful within the group at t_0, the more likely it is that ego perceives the focal actor as powerful at t_1. That is, to deal with uncertainty concerning who is powerful, individuals tend to rely on other actor's perceptions of the focal actor's informal power. This finding is in line with previous studies reporting individual tendencies to perceive popular actors as even more popular than they really are (Kilduff, Crossland, Tsai, & Krackhardt, 2005). In a similar way a large number of power attributions to a focal actor within a group increases the likelihood that the focal actor will be perceived by ego as powerful or perhaps even more powerful over time. More generally, the findings underline the importance of imitation as a strategy to cope with uncertainty concerning social processes in groups.

Further, we found support for a social embeddedness argument (the interpersonal ties hypothesis). Close interpersonal trust relationships between actors serve as a source of information concerning the focal actors' interactions with other group members, as well as his or her degree of influence in such relationships. Our finding supports the idea that individuals tend to rely on the trusted person's accounts when attributing power to him/her. Therefore, the more ego trusts the focal actor, the more likely he is to perceive the focal actor as powerful over time. This result indicates the importance of considering egos' personal ties to other organizational members as a valuable source of information on (changes in) their relative power position.

As far as the effect of power strategies is concerned, our hypotheses found only partial and weak support. The direct strategic behavior has no effect on individual's

informal power. A possible explanation for this finding might be that this type of strategic behavior may be visible only to a very limited number of the organizational actors, and hence less relevant for the change in one's power reputation in the group as a whole. Furthermore, we found no support for our hypothesis concerning the effect of indirect strategic behavior on perceived informal power dynamics. A possible explanation for this result could be related to measurement issues. For this study we chose to build on the distinction between direct and indirect power strategies (Morrill, & King Thomas, 1992). It is possible that the chosen measure could not capture the strategic behavior characteristic of actors in our sample. Future research might eventually benefit using other measures of power strategy use which have been proposed in the literature and focus on other theoretical dimensions (Kellerman & Cole, 1994). For example, focusing on managerial power strategies, Gioa and Sims (1983) suggest distinguishing between positive, punitive, and goal-setting strategies (rather than building on the distinction between direct and indirect strategies) as behaviors that will differentially affect power perceptions of subordinates. Furthermore, the measure is based on self-reports and focuses on the perceived appropriateness of certain strategies rather than actual strategy use. Our results suggest that to measure strategic behavior of organizational actors it might be important for future research to incorporate different reporters of the *actual* strategy use in organizational setting.

We found support for our hypothesis concerning the negative effect of (indirect) formal strategy use on perceived informal power dynamics. Organizational members appear to interpret this type of strategic actions by a focal actor as a signal of his inability to get things done. Individuals activating formal authorities to influence others are therefore less likely to be viewed as informally powerful by other group members. This finding is in line with earlier research suggesting that connections with high-ranking supervisors could detract from one's reputation rather than enhance it (e.g., Mehra, Dixon, Brass, & Robertson, 2006).

Our results indicate that in addition to the effects of the individual level variables, the perceived power dynamics are strongly affected by what's going on in the organizational setting throughout the measurement period. In the setting under study, power play in the wider organizational context not only resulted in a temporary decrease of the site manager's informal power, but led to an overall depletion of informal power attributions in the whole system. This aspect of power dynamics in

the network is in line with ethnographic observations which showed a general decrease of communication and interpersonal trust during this phase (Wittek, 1999).

On a theoretical level, these processes lend strong support to earlier suggestions indicating the need to consider transfer of power across organizational levels (Fiol et al., 2001), and the importance to take context effects seriously (Johns, 2006).

The following limitations of the present study should be noticed. First, the current results build upon one single case, the management team of a German Paper Factory, consisting of relatively few persons. Hence, more research on different organizational settings and contexts is needed to be able to generalize our results. Second, our operationalization of power strategies focuses on tactics that might be less relevant for capturing strategic behavior that others perceive as cues for an individual's power. Finally, whereas our power perceptions were measured at four points in time, power strategies were measured only at two points in time (i.e., in wave one and wave three, respectively), and therefore wave two and four had to be imputed by values obtained from the last available measurement point. The imputation procedure could have an effect on the obtained results. Hence, future studies could benefit form a better measurement of strategic behavior.

Despite the mentioned limitations, this study represents an important contribution to the existing empirical research focused on modeling the antecedents and dynamics of perceived differences in informal power reputation. Empirical studies on perceived power dynamics in real life settings, based on longitudinal network data, are scarce. An important conclusion from our results is that stability or change of perceptions concerning someone's informal power position are driven by other actor's perceptions of the focal actor's informal power, the focal actor's social embeddedness in networks of interpersonal relationships, and the focal actor's use of power strategies. Questions raised by the current study indicate that further research on informal power perceptions has the potential to offer additional fruitful and necessary insights concerning the antecedents and dynamics of perceived differences in informal power reputation.

Where does Reputational Power in Organizations Come From

References

Aghion, P. & Tirole, J. (1997). Formal and real authority in organizations. *The Journal of Political Economy, 105,* 1-29.

Bacharach, S.B., &. Lawler, E.J. (1976). The perception of power. *Social Forces,* 55, 121-134.

Barrera, D., & Buskens, V. (2007). Imitation and learning under uncertainty – A vignette experiment. *International Sociology,* 22 (3), 367-396.

Boer, P., Huisman M., Snijders, T. A. B., Steglich, C. E. G., Wichers, L. H. Y., Zeggelink, E. P. H. (2006). *StOCNET: an open software system for the advanced statistical analyses of social networks. Version 1.7.,* Groningen, ICS/SciencePlus.

Brass, D. J. (1984). Being in the right place: A structural analysis of individual influence in an organization. *Administrative Science Quarterly, 29,* 518-539.

Brass, D. J. & Burkhardt, M. E. (1993). Potential power and power use: An investigation of structure and behavior. *The Academy of Management Journal, 36,* 441-470.

Cross, R. & Prusak, L. (2002). The people who make organizations go-or-stop. *Harvard Business Review, 80,* 104-112.

DiMaggio, P., & Powell, W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review,* 48 (2), 264/

Fiol, M., O'Connor, E. J., & Aguinis H.. (2001). All for one and one for all? The development and transfer of power across organizational levels. *Academy of Management Review,* 26 (2), 224-242.

Friedkin, N. (1983). Horizons of observability and limits of informal control in organizations. *Social Forces.*

Gioa, D., & Sims, H. Jr. (1983). Perceptions of managerial power as a consequence of managerial behavior and reputation. *Journal of Management,* 9 (1), 7-26.

Gould, R. (2002). "The origins of status hierarchies: A formal theory and empirical test." *American Journal of Sociology,* 107:1143-78.

Hedstrom, P. (1998). Rational Imitation. Pp. 306-27 in *Social Mechanisms: An Analytical Approach to Social Theory,* edited by. P. Hedström and R. Swedberg. Cambridge: Cambridge University Press.

Hess, N., & Hagen H. (2006). Psychological adaptations for assessing gossip veracity. *Human Nature,* 17 (3), 337-354.

Hinkin, T., & Schriesheim, C. (1994). An examination of subordinate-perceived relationships between leader reward and punishment behavior and leader bases of power. *Human Relations,* 47 (7), 779-800.

Ibarra, H. (1993). Network centrality, power, and innovation involvement: Determinants of technical and administrative roles. *The Academy of Management Journal, 36,* 471-501.

Johns, G. (2006). The essential impact of context on organizational behavior. *Academy of Management Review, 31*, 386-408.

Kellermann, K., & Cole, T. (1994). Classifying compliance gaining messages: Taxonomic disorder and strategic confusion. *Communication Theory, 4 (1),* 3-60.

Kilduff, M. & Krackhardt, D. (1994). Bringing the individual back in: A structural analysis of the internal market for reputation in organizations. *The Academy of Management Journal, 37,* 87-108.

Kilduff, M., Crossland, C., Tsai, W., & Krackhardt, D. (2005). A small world after all? Network perceptions versus reality. Paper presented at the Academy of Management meetings. Hawaii.

Krackhardt, D. (1990). Assessing the political landscape: Structure, cognition, and power in organizations. *Administrative Science Quarterly, 35,* 342-369.

Krackhardt, D. & Hanson, J. R. (1993). Informal networks: The company behind the chart. *Harvard Business Review, 71,* 104-111.

Kumbasar, E., Romney, K. A., & Batchelder, W. H. (1994). Systematic biases in social perception. *American Journal of Sociology, 100,* 477-505.

Kyl-Heku, L. M. & Buss, D. M. (1996). Tactics as units of analysis in personality psychology: An illustration using tactics of hierarchy negotiation. *Personality and Individual Differences, 21,* 497-517.

Lund, O. C. H., Tamnes, C. K., Moestue, C., Buss, D. M., & Vollrath, M. (2007). Tactics of hierarchy negotiation. *Journal of Research in Personality, 41,* 25-44.

Mehra, A., Dixon, A. L., Brass, D. J., & Robertson, B. (2006). The social network ties of group leaders: Implications for group performance and leader reputation. *Organization Science, 17,* 64-79.

Meindl, J. R., Ehrlich, S., & Dukerich, J. M. (1985). The romance of leadership. *Administrative Science Quarterly, 30*, 78-102.

Milliken, F. (1987). Three types of perceived uncertainty about the environment: State, effect, and response uncertainty. *Academy of Management Review, 12,* 133-143.

Where does Reputational Power in Organizations Come From

Morrill, C., & King Thomas, C. (1992). Organizational conflict management as disputing process. The problem of social escalation. *Human Communication Research, 18 (3)*, 400-428.

Pastor, J. C., Meindl, J. R., Mayo, M. (2002). A network effects model of charisma attributions. *Academy of Management Journal, 45*, 410-421.

Pfeffer, J. (1977). The ambiguity of leadership. *Academy of Management Review, 2*, 104-112.

Pfeffer, J. & Fong, C. T. (2005). Building organization theory from first principles: The self-enhancement motive and understanding power and influence. *Organization Science, 16,* 372-388.

Roberts, N. (1986). Organizational power styles: Collective and competitive power under varying organizational conditions. *Journal of Applied Behavioral Science, 22,* 443-458.

Sharman, J. C. (2007). Rationalist and constructivist perspectives on reputation. *Political Studies, 55,* 20-37.

Snijders, T. A. B. (2005). "Models for longitudinal network data" in Carrington, P., Scott, J., & Wasserman, S. (eds.), *Models and methods in social network analysis*, New York, Cambridge University Press.

Snijders, T. A. B., Steglich, C. E. G., Schweinberger, M., & Huisman, M. (2007). Manual for SIENA version 3.1 – print version (version 16 April 2007). Available at: http://stat.gamma.rug.nl/stocnet/

Weary, G., & Edwards, J. (1994). Individual differences in causal uncertainty." *Journal of Personality and Social Psychology, 67 (2),* 308-318.

Wittek, R. (1999). *Interdependence and informal control in organizations.* University of Groningen, Amsterdam: Thela Thesis.

Wittek, R. (2003). Mimetic trust and intra-organizational network dynamics. *Journal of Mathematical Sociology,* 25 (1), 109-138.

Where does Reputational Power in Organizations Come From

Tables

Table 1. *Descriptive Statistics of Actors (N=17). Average Degree for Influence and Trust Networks and Means for Power Strategies.*

Network characteristic	Time point 1	Time point 2	Time point 3	Time point 4
Average degree influence	1.69	4.56	1.19	5.44
Average degree trust	1.75	5.13	1.06	3.00
Bilateral arguing	0.49	0.49	0.29	0.29
Public negotiation	0.60	0.60	0.32	0.32
Retaliation	-0.92	-0.92	-0.88	-0.88
Resignation	-0.72	-0.72	-0.62	-0.62
Complain to superiors	-0.28	-0.28	-0.22	-0.22

Where does Reputational Power in Organizations Come From

Table 2. *Informal Power over Time. Parameter Estimates and Standard Errors (SE) for Network Dynamics.*

	Model	
Control variables	Estimate	SE
1. Network rate parameter (period 1)	8.67	1.85
2. Network rate parameter (period 2)	6.97	1.13
3. Network rate parameter (period 3)	25.88	10.02
4. Outdegree (density)	-3.55	0.40***
5. Dummy period 2	-2.11	0.54***
Structural effects		
6. Popularity of alter	0.66	0.11***
Dyadic attributes / explanatory network variables		
7. Trust	0.40	0.08***
8. Formal position	0.35	0.29
Actor attributes (direct strategy)		
9. Bilateral arguing	-0.07	0.14
10. Public negotiation	0.26	0.22
Actor attributes (indirect strategy)		
11. Retaliation	-1.08	0.65 ^{o}
12. Resignation	0.22	0.14
Actor attributes (indirect formal strategy)		
13. Complain to superiors	-0.30	0.14 *

Note: *** p<0.001; ** p<0.01; * p<0.05; ^{o} p<0.10.

Why Bother with What Others Tell You? An Experimental Data-Driven Agent-Based Model

Riccardo Boero[*], Giangiacomo Bravo[†], Marco Castellani[°], Flaminio Squazzoni[°]

[*] Dipartimento di Scienze Economiche e Finanziarie "G. Prato", Università di Torino and GECS — Research Group in Experimental and Computational Sociology, Università di Brescia
boero@econ.unito.it
[†] Dipartimento di Scienze Sociali, Università di Torino and GECS — Research Group in Experimental and Computational Sociology, Università di Brescia
giangiacomo.bravo@unito.it
[°] Dipartimento di Studi Sociali and GECS — Research Group in Experimental and Computational Sociology, Università di Brescia {castella, squazzon}@eco.unibs.it

Abstract. This paper investigates the relevance of reputation for enhancing the exploration capabilities of agents in uncertain environments. We present a laboratory experiment where 64 subjects are called to take iterated decisions on economic investment. Behavioral patterns followed by subjects are first traced and recognized and then validated through an agent-based model that exactly replicates the experiment and the recognized patterns. Once confirmed the experimental patterns, we have created a experimental-data driven agent-based model to investigate the effect of reputation on the exploration capabilities at the system level. The results show that reputation significantly increases the capability of agents to cope with uncertain environments.

Keywords: reputation; exploration vs. exploitation; laboratory experiments; agent-based models; social simulation.

1 Introduction

In the real life, agents do not decide what to do just relying on their calculating capabilities but often communicate with other qualified people before deciding what to do, and are influenced by what others think and tell as well. This happens, for instance, when private estate owners are searching for efficient builders to restore mansions in a particular area to increase revenues on a real estate market, or when entrepreneurs try to sort out trustworthy companies to subcontract a production phase. This is a well-known situation when agents behave on markets characterised by uncertainty, information asymmetries and ambiguity. In such a situation, quality of goods exchanged is difficult to evaluate and guarantee, moral hazard among parties involved in a transaction can take place and market failures can easily result (e.g., Akerlof 1970; Williamson 1979). The consequence of this is that agents are likely to be much sensitive to gossip, reputation and other social information sources (Conte

and Paolucci 2002; Sommerfeld *et al*. 2007). This is where the social dimension of the economic action enters the picture (Granovetter 1974; Raub and Weesie 1990; e.g., the case of financial markets in Callon 1998; Beunza and Stark 2003; Knorr Cetina and Preda 2004; Burt 2005). Such crucial determinants of the concrete social contexts where economic action takes place have been investigated also in some recent studies in economics (e.g., Durlauf and Young-Peyton 2001; Gui and Sugden 2005).

Information exchange and signals can help reduce the gap between what agents know and what they should know and take the risk of entering in interaction with others. The intriguing issue is that, particularly on markets, information, e.g. what agents know about hot and strategic issues, is often kept secret or distorted to gain profit. This happens because information is a relevant asset and can be negotiated, exchanged or protected (Burt 2005). When this happens, the problem for agents is to detect cheaters and trustworthy information sources so to identify the reliability of information.

Following this theoretical background, the paper investigates how reputation and social information can affect, at the micro level, the performance of economic agents in uncertain environments and, at the macro level, the exploration capability exhibited by the system as a whole. In particular, we focus on the capability of agents to detect trustworthy information sources while dealing with risky investments and exploration processes. In this perspective, we compare social systems where agents can rely just on their personal experience and where agents can rely on reputation mechanisms.

The method suggested in this paper is a combination of laboratory experiments and agent-based social simulation (e.g., Janssen and Ostrom 2006). The first step has been to design a laboratory experiment to generate robust data on behaviour of human subjects in a well controlled decision setting (Boero and Squazzoni 2005). To capture the widest possible range of individual decision in uncertain environments, we have created an *ad hoc* experiment that is similar to an external observation of human decision in a controlled decision setting. This is, firstly, to avoid to underpin our experiment on a pre-established standard framework and to be able to trace the highest heterogeneity of subjects' behaviour. Secondly, this is to explicitly design an experiment where the explanatory purpose is not to understand the impact of a treatment on a particular variable, neither to analyse the systemic consequence of the treatment itself, but rather to identify the behaviour of human subjects up against an uncertain decision environment in a bottom-up fashion. The second step has been the analysis of the behavioural patterns followed by human subjects, called pattern recognition and classification. We have used a two-step cluster analysis on experimental data to identify behavioural patterns. The third step has been to design an agent-based model to validate the patterns identified in the experimental data, by creating a model that rerun the experiment and replicated the behavioural patterns previously identified. The assumption is that whether we were able to exactly replicate the experimental results this should meant that the abstraction of the patterns imposed to experimental data did not imply any information loss. The next step has been to design an experimental evidence-based social simulation to introduce interaction among agents so to explore the impact of some relevant simulation parameters on the micro-macro outcomes.

2

The paper is organised as follows. In the second section we describe the experiment. In the third one we summarise the experimental results. By applying a two-step cluster analysis to the data, we have identified some clear-cut patterns of behaviour followed by the subjects. In the fourth section we describe the process of validation of the experimental patterns through an agent-based model that exactly replicated the experiment and the behavioural patterns previously identified. In the fifth section we introduce an experimental-data driven agent-based model to understand the relevance of agents' interaction and of reputation mechanism in generating macro outcomes.

In conclusion, this paper presents a cross-methodological exercise that shows how much explanatory power can be achieved when agent-based models are based on well controlled empirical/experimental data. The evidence is twofold: first, we show that human subjects tend to exhibit a certain degree of coherence in different decision domains; secondly, we show that reputation mechanisms allow people to generate more efficient exploration patterns in uncertain environments, although an amount of false information circulates within the system.

2 The "Find the Best" Experiment

Sixty four subjects participated to the experiment that took place between October and November 2007 in two days in the computer laboratory of the Faculty of Economics of the University of Brescia, which is equipped with the experimental software z-Tree (Fischbacher 1999). Subjects were students belonging to different faculties of the University of Brescia, 38 females and 26 males, recruited through public announcements. The subjects played 17 rounds of the FTB "Find the Best" game. Players have an initial endowment of 1000 ECU and a randomly assigned security with a given yield. The space consists of 30 securities with unknown yield. Players can ask to have a new security each round. They know just that the game is supposed to end with a probability = 0.1. In each round players know just the securities discovered in the previous rounds. To discover a new security players have an exploration cost of 100 ECU. Whenever players decide not to explore a new security, at the end of the round their wealth follows this function: $w = e_i + M_s$ where e_i is the endowment of the previous round and M_s the maximum-yield security discovered by players. Whenever players decide to explore a new security, their wealth is as follows: $w = (e_i - explo) + P_s$ where e_i is the initial endowment, *explo* is the cost of exploration and P_s is the yield of the new security discovered through the exploration. The expected profit of players follows this function: $p = w + (M_s \times 10)$, where w is the wealth and M_s is the maximum-yield security discovered by players. Players know that expected profit will be paid at the end of the game.

For example, player *A* has an initial endowment of 1000 ECU and assigned a security that yields 20 ECU. Let us suppose that, at the first round, *A* chooses to explore the space, asks for another security and discovers a security with 30 ECU of yield. At the end of the first round *A* would have a wealth of 930 ECU = 1000 ECU (e_i) -100 ECU (*explo*) + 30 ECU (P_s), but an expected profit of 1230 ECU = 930 ECU (w) + (30 ECU x 10) if the first round were the end of the game. Let us suppose

that *A* at the second round chooses to further explore a new security and discovers a security with 25 ECU of yield. At the end of the second round, *A* would have a wealth of 855 ECU = 930 ECU (e_i) -100 ECU (*explo*) + 25 ECU (P_s), but an expected profit of 1155 ECU = 855 ECU (*w*) + (30 ECU x 10), since M_s, that is, the maximum- yield security discovered by players, is still the 30 ECU security. This is to introduce risk investment.

The order of securities' exploration is a fixed sequence of yields as follows: 27, 23, 30, 30, 28, 39, 46, 53, 48, 58, 63, 57, 64, 61, 59, 66, 69, and 72 ECU over 17 rounds. This means that whenever players ask for a new security will receive a 27 yield security in the first round, a 23 yield security in the second round, a 20 yield security in the third one, and so on until round 17.

Players are fictiously divided in groups of four players. Each player supposes to repeatedly interact with other three players, fictiously called Tom, Dick and Harry, which in reality are automata with a fixed behaviour. Each round, the decision of each player is composed of two parts: (1) what to do with securities (exploit the already discovered security or explore a new security?) and (2) what information on the securities already discovered transmit to Tom, Dick and Harry. Information can cover just already discovered securities and can be true or false. Player *A* can decide between four kind of information:

i) *A* makes known to Tom, Dick and Harry that a given security (that is M_s that is the maximum-yield security discovered by *A*) yields exactly what it yields (true information), even if Tom, Dick and Harry do not know that the security in question is the M_s of player *A*;
ii) *A* makes known to Tom, Dick and Harry that a given security, namely an average yield security discovered by *A*, yields exactly what it yields (true information);
iii) *A* makes known to Tom, Dick and Harry that a given security (that is M_s that is the maximum-yield security discovered by *A*) eventually yields little (false information);
iv) *A* makes known to Tom, Dick and Harry that a given security, namely an average yield security discovered by *A*, yields more that what it really yields (false information).

This holds for the first five rounds. The meaning of *A* sending information to Tom, Dick and Harry without any consequence or feedback is to induce *A* to believe that there is an interaction among players. During the subsequent 12 rounds, players *A* begin to receive hints on security yields from Tom, Dick or Harry as informers. Informers follow three fixed attitudes:

A= they communicate hints that are always true;
B= hints concern securities that yield less than what is communicated;
C= they communicate hints that are always false because yields are randomly higher or lower than what they really are.

By introducing Tom, Dick and Harry as group informers we dramatically change the possibility to explore the space of securities for players. Now exploration is not

just driven by random search, as in the first five rounds, but also by the availability and possibility to use hints from other people who are supposed (wrongfully or rightfully) to know more. The process of information transmission is fixed, as well as what the informers respectively do for each round. As shown in Table 1 whenever players *A* decide to use information coming the informers, so as to explore new securities, there is fixed outcome for each round: at round 6, hints on securities will be true (=A), while at round 7 they will be false (=B), and so forth.

Round	Information sent to Players	Effective Yield	Informers A= always the truth B= always higher yields C=higher or lower yields
6	48	48	A
7	52	44	B
8	56	46	B
9	50	48	C
10	62	62	A
11	70	70	A
12	74	58	B
13	73	75	C
14	80	66	C
15	88	88	A
16	94	96	C
17	96	68	B

Table 1. Structure of informers' hints from round 6 to the end of the game.

Now, the decision of each player is composed of three steps. The first is what to do with securities. On this, there are the following options:

1) exploiting M_s, that is, the maximum-yield security discovered (available also in the first five rounds);
2) exploring the space by discovering a new security via random search (available also in the first five rounds);
3) exploring the space by following informers' hints (introduced since round 6).

The second step is the same as in 1-5 rounds and concerns the information transmission from *A* to the fictious group members, Tom, Dick and Harry. The options are the four ones mentioned above (i; ii; iii; iv). The information from *A* is addressed to the same member (Tom, Dick or Harry) who has transmitted the information to *A* at the first step.

To conclude, the "FTB game" issues two challenges to players: maximise their expected profit at the end of the game by exploiting or exploring the space of possibilities, and discovering/understanding when the hint of the informers is true and follow this.

3 Experimental Results: Pattern Recognition

By means of a cluster analysis it is possible to recognize and abstract away some behavioural patterns on experimental data, both those concerned with information transmission and with decision on securities. The cluster analysis has been applied to data on players' decision as regards to information transmission. The analysis dissected three clusters called I1, 2, and 3. They are as follows:

i) I1 cluster (26,6% of subjects) is characterised by cooperative players who transmitted true information regardless of the quality of information they have received by the informers;
ii) I2 cluster (50% of subjects) is characterised by conditionally cooperative players who tended to transmit true information but also to reciprocate false information in turn;
iii) I3 cluster (23,4% of subjects) is characterised by cheaters who mostly exaggerated the yields of the discovered securities regardless of the quality of information they have received by the informers.

As shown in Figure 1, while in about 80% of the cases I1 players told the truth by transmitting to others the exact yield of the best security discovered, in about 80% of the cases I3 players told the false, half the cases transmitting to others higher yields of the security discovered. I2 players told the truth in 75% of the cases, transmitting to others the exact yield of the best security discovered or other high yields.

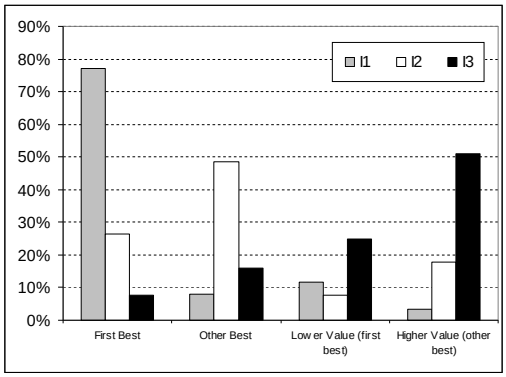


Figure 1. Information transmitted by players divided into three different clusters, L1, I1, and I3.

Table 2, 3, and 4 show data that help confirm the plausibility of the results of the cluster analysis. Table 2 shows that I1 players told the truth both when they met trustworthy and false informers (A= 88%; C = 88%), I2 players told almost every time the truth but were more prompt than I1 players to reciprocate with the informers (35% false information to B and 25% false information to C informers), whereas I3 players transmitted false information in 76% of the cases. Table 3 shows that I1 players told the truth by transmitting reliable information on best discovered yields (91%)), while I2 and I3 players preferred to exaggerate the yield of the discovered

securities (65% for I2 and 67% for I3 players). Table 4 shows that I1 players, when decided to transmit false information, preferred to tell that yields had lower values than higher ones, I2 and I3 players told a partial truth, by transmitting not the best yield discovered but some second-best ones. This means that, when I1 players decided to cheat others, they did not played it hard so as to not take high responsibility in inducing others in following the false information they decided to transmit.

The evidence is that the behavioural clusters show a certain coherence, a small degree of conditionality of the cooperation can be found in all the clusters, but with significantly lower impact on I1 and I3, and that subjects have perfectly understood the game. There is not noisy or illogical behaviour, and players have perfectly understood the different trustworthiness of A, B, and C hints.

Informers	I1Players who returned true info	I1 Players who returned false info	I2 Players who returned true info	I1 Players who returned false info	I1 Players who returned true info	I1 Players who returned false info
A	88%	12%	88%	12%	27%	73%
B	78%	22%	65%	35%	20%	80%
C	88%	12%	71%	29%	25%	75%
tot	85%	15%	75%	25%	24%	76%

Table 2. Percentage of players who returned true/false information to informers divided into three clusters, I1, I2, and I3.

Informers	I1 Players who returned first best info	I1 Players who returned other best info	I1 Players who returned first best info	I2 Players who returned other best info	I1 Players who returned first best info	I3 Players who returned other best info
A	92%	8%	38%	62%	6%	94%
B	89%	11%	35%	65%	33%	67%
C	92%	8%	32%	68%	60%	40%
tot	91%	9%	35%	65%	33%	67%

Table 3. Type of information returned by players who returned true information to informers divided into three clusters, I1, I2, and I3.

Informers	I1 Players who suggested that securities have lower value (first best)	I1 Players who suggested that securities have higher values (other best)	I2 Players who suggested that securities have lower value (first best)	I2 Players who suggested that securities have higher values (other best)	I3 Players who suggested that securities have lower value (first best)	I3 Players who suggested that securities have higher values (other best)
A	100%	0%	47%	53%	50%	50%
B	67%	33%	33%	67%	27%	73%
C	75%	25%	19%	81%	22%	78%
tot	77%	23%	30%	70%	33%	67%

Table 4. Type of information returned by players who returned false information to informers.

The next step has been to extract some simplified behavioural patterns that have been translated in a computational pseudo-code that has been used to build the agent-

based model that replicated the experiment. Table 5, 6 and 7 show the pseudo-code for I1, I2, and I3 players.

```
if (p = 0.75)
     FIRST_BEST;
else {
     if (p = 0.75) LOWER_HIGH;
     else HIGHER_LOW;
}
```

Table 5. The pseudo-code to trace the behaviour of I 1 players.

```
if (partner trustworthy){
     if (p = 0.25) FIRST_BEST;
     else OTHER_BEST;
} else {
     if (p = 0.75) {
           if (p = 0.25) FIRST_BEST;
           else OTHER_BEST;
     } else {
           if (p = 0.25) LOWER_HIGH;
           else HIGHER_LOW;
     }
}
```

Table 6. The pseudo-code to trace the behaviour of I2 players.

```
if (p = 0.25){
     if (p = 0.25) FIRST_BEST;
     else OTHER_BEST;
} else {
     if (p = 0.25) LOWER_HIGH;
     else HIGHER_LOW;
}
```

Table 7. The pseudo-code to trace the behaviour of I3 players.

The next step has been to apply the same cluster analysis to data on decisions on securities. As mentioned before, in the experiment, players were called to decide between random exploration of new securities, hint-following exploration of new securities, and exploitation of already discovered securities. The results of the cluster analysis allows us to dissect three clusters called A1, A2, and A3 as follows:

i) A1 cluster (31,3% of subjects) is characterised by explorative players who took the risk of exploring the space of possibility following the informers' hints;

ii) A2 cluster (14,1% of subjects) is characterised by players who decided depending on the reliability of informers; when the informer is trustworthy, they decided to explore but often following random exploration; when the informer is not trustworthy, if they decided to explore the space, they did not follow informers' hints;

iii) A3 cluster (54,7% of subjects) is characterised by players who followed the informers' hints when informers are reliable, whereas they decided to exploit already discovered securities whether the informers hints were not reliable.

Figure 2 summarises the distribution of strategies followed by players. As said before, A1 players preferred to explore the space (75% of cases), in most cases by following informers' hints, A2 players did not explore the space so much (55% of cases) but when they did it they decided to follow a random exploration, and A3 players preferred not to explore (55% of cases) but when they did it they followed informers' hints.

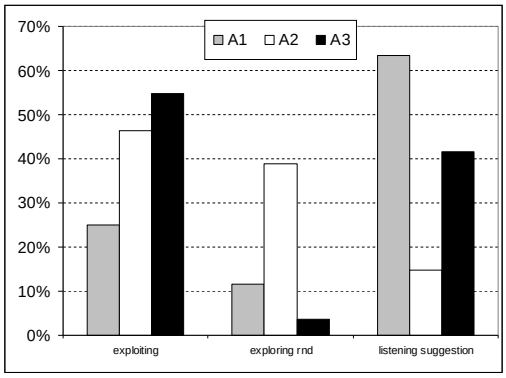


Figure 2. Strategies of space exploration followed by players divided into the three clusters, A1, A2, and A3.

Table 8 and 9 allow us to confirm the results of the cluster analysis. Table 8 shows that A1 players preferred to explore regardless of informers' hints (there is no difference when hints come from A, B, or C), A2 players are sensitive to reliability of informers (they followed true A hints in 72% of case, whereas followed C false hints in 44% of cases), and A3 players followed the evaluation of the reliability of informers in 2/3 of the cases. The behaviour of A2 players is statistically a minority but is really interesting. As it is shown in Table 9, such players were sensitive to the reliability of the informers' hints but they decided to randomly explore the space, as though they was viewing hints as mere signals of the availability of potentially new yields. In a certain sense, they separated the problem of the reliability of informers from the quality of the information.

Like in the previous case, the next step has been to extract some simplified behavioural patterns that have been translated in a computational pseudo-code that has been used to build the agent-based model that replicated the experiment. Table 10, 11 and 12 show the pseudo-code for A1, A2, and A3 players.

Informers	A1 Players who decided to explore new securities	A1 Players who decided to exploit old securities	A2 Players who decided to explore new securities	A2 Players who decided to exploit old securities	A3 Players who decided to explore new securities	A3 Players who decided to exploit old securities
A	79%	21%	72%	28%	69%	31%
B	73%	28%	44%	56%	36%	64%
C	74%	26%	44%	56%	30%	70%
tot	75%	25%	54%	46%	45%	55%

Table 8. Percentage of players who followed exploration vs. exploitation divided into the three clusters, A1, A2, and A3.

Informers	A1 Players who explored randomly	A1 Players who followed informers' hints	A2 Players who explored randomly	A2 Players who followed informers' hints	A3 Players who explored randomly	A3 Players who followed informers' hints
A	6%	94%	77%	23%	7%	93%
B	29%	71%	88%	13%	10%	90%
C	12%	88%	50%	50%	7%	93%
tot	16%	84%	72%	28%	8%	92%

Table 9. Percentage of explorative players who followed random exploration or informers' hints, divided into the three clusters, A1, A2, and A3.

```
if (p = 0.75) {
     if (p = 0.25) EXPLORING;
     else LISTENING;
} else     EXPLOITING;
```

Table 10. The pseudo-code to trace the behaviour of A1 players.

```
if (partner trustworthy){
     if (P = 0.75) {
            if (p = 0.75) EXPLORING;
            else LISTENING;
     } else EXPLOITING;
} else {
     if (p = 0.5) {
            if (p = 0.75) EXPLORING;
            else LISTENING;
     } else EXPLOITING;
}
```

Table 11. The pseudo-code to trace the behaviour of A2 players.

```
if (partner trustworthy){
     if (p = 0.75) LISTENING;
     else EXPLOITING;
} else {
     if (p = 0.25) LISTENING;
     else EXPLOITING;
```

10

```
}
```

Table 12. The pseudo-code to trace the behaviour of A3 players.

The next step has been to combine the two cluster analyses so to recognize whether players would have shown a coherence between decisions in different domains, such as information and space exploration. Our hypothesis was that such decisions, although being oriented to different and hypothetically separated decision domains, would not being perceived by players as totally independent.

Table 13 shows two sound evidences on behavioural patterns, at the same time providing a sound confirmation of the hypothesis mentioned above. First, **players who did not follow informers' hints suggested in turn false information to others**. The evidence is that A2 players, who decided to explore the space randomly, that is, without following informers' hints, are the same I3 players who decided to transmit false information to others. A2-I3 players show a particular and coherent behavioural pattern, i.e., they have decided to explore the space randomly, when they knew from informers that other securities were available, but they did not trust informers deciding at the end to follow random exploration. When they are called to send information, they consequently sent false information to others. Secondly, **players who followed informers' hints suggested in turn true information to others**. In fact, table 13 shows that A1 and A3 players are especially distributed in I1 and I2 clusters.

	A1	A2	A3
I1	40%	11%	23%
I2	55%	33%	51%
I3	5%	56%	26%

Table 13. Distribution of players for I clusters according to A clusters belonging.

4 Validation of Experimental Patterns

The next step has been to validate the patterns that we have previously identified on experimental data through an agent-based model that aimed to replicate exactly the experimental conditions. All the simulation parameters have been settled accordingly. The agent-based model has been based on behavioural algorithms that were aimed to specifically reproduce the behavioural patterns described in the third section. Agents are 64, are matched with A, B, and C, that is, the informers, which follow the same fixed behaviour they followed in the experiment. As in the experiment, at each run of the simulation agents are called to take two decisions: action (i.e., random exploration, hints' driven exploration or exploitation) and information (i.e., first best, other best, lower value or higher value). Since the model makes use of behavioural algorithms with probabilistic features, we have run 1000 simulations of the same model and averaged the results with different random number generators.

Simulation results have largely replicated the behavioural patterns recognized in the experiment and all the stylized facts that we could abstract away from the

experimental evidence. Table 14 summarises the comparison of experimental and simulated data and shows that average value and standard deviation at the end of the game/simulation are quite similar. Figure 3 shows a comparison of the distribution of final profit at the end of the game/simulation. Looking at the Gaussian, the evidence is similar, although at the end simulation data showed a higher degree of heterogeneity.

A stronger evidence of the success of the replication came from the distribution of players/agents' decisions. Figure 4 shows a comparison of the strategies that players/agents have followed. The distribution is quite similar. As it is shown in Figure 5, the same holds true for what players/agents did with information.

	Experiment		**Simulation**	
	Average	Standard Deviation	Average	Standard Deviation
Endowment	981,88	292,932	944,6487	275,72832
BestDiscovered	88,03	11,506	85,4917	5,57511
FinalProfit	1862,19	309,881	1799,5653	286,87753

Table 14. Comparison of experimental and simulated results. *Endowment* refers to the average amount of assets cumulated over time by the players. *BestDiscovered* refers to the best yield discovered by players. *FinalProfit* is the profit achieved at the end of the game/simulation.

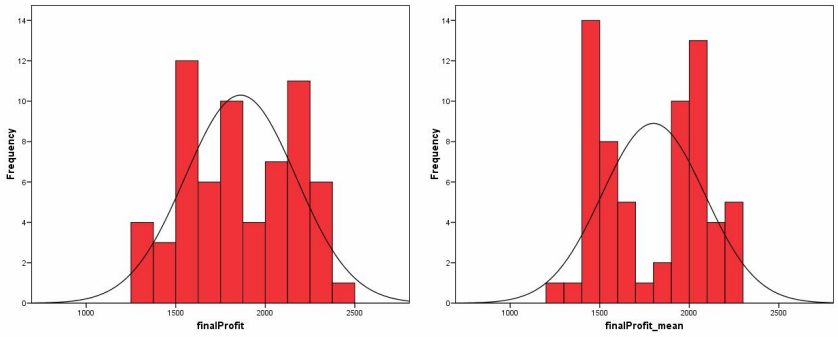


Figure 3. Distribution of final profit achieved by players/agents. On the left the experimental results. On the right the simulation results.

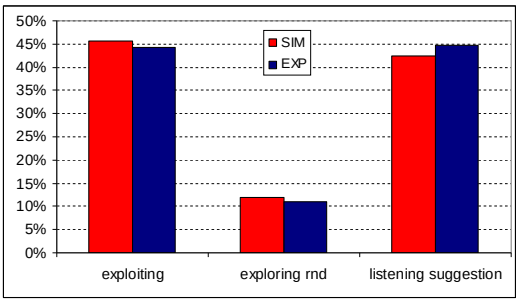


Figure 4.Comparison of players'/agents' action. Simulation results are red, experimental results are blue.

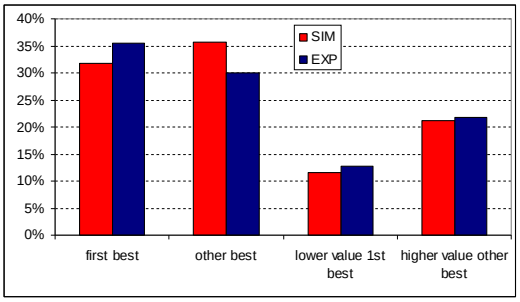


Figure 5.Comparison of players'/agents' decision on information. Simulation results are red, experimental results are blue.

The validity of the replication is further corroborated when we have a look at the dynamics of the factors previously taken into account. In figures 6, 7 and 8 we show the comparison of dynamics of *final profit*, *endowment*, and *best security discovered* between experimental and simulation data. As one can see, there are quite similar dynamics. A further confirmation can be found if we look at the Pearson correlation coefficient of the representative variables: "final profit" (0,99964), "endowment" (0,971084), and "best discovered" (0,999154).

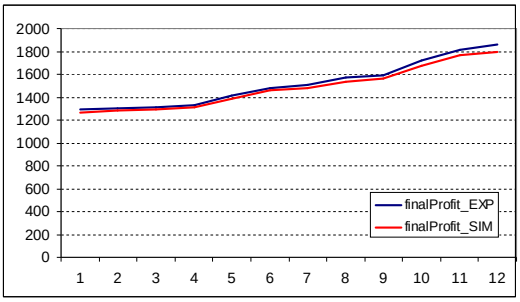


Figure 6. Dynamics of *final profit* over time in the experiment and in the simulation.

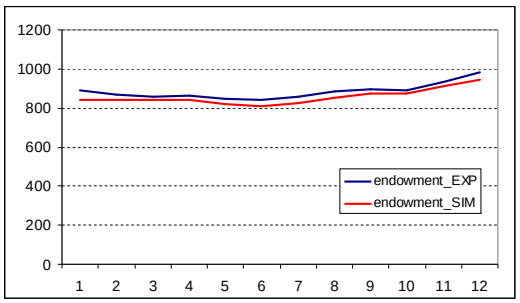


Figure 7. Dynamics of *endowment* over time in the experiment and in the simulation.

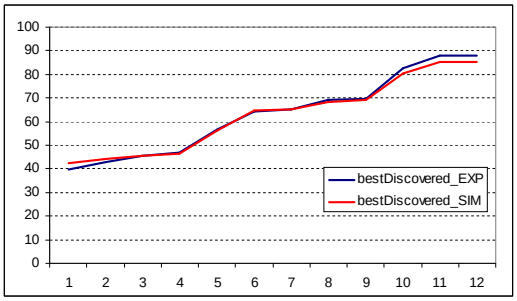


Figure 8. Dynamics of *best discovered security* over time in the experiment and in the simulation.

In conclusion, the simulation exercise unequivocally demonstrates that the analytic abstraction that we have pursued by identifying, recognizing and synthesizing behavioural patterns on experimental data did not mean a significant information loss as regards to what really happened in the experiment. This step allowed us to have a well controlled experimental evidence on which to build further simulations that were aimed to deal with the analytical purpose mentioned in the introduction.

5 The Agent-Based Model

The next step has been to build an agent-based model based on the computationally validated experimental patterns described in the previous section so to explore some first research questions on the impact of reputation on exploration of space of possibility at the agent and at the system level. As we show in Table 15, we have exactly followed the distribution of experimental patterns in designing the agent-based model. The model consists of 100 agents that should improve their profit by moving across an uncertain environment of security yields, like in the experiment. “Endowment” and “final profit” follow the same function of the experiment.

The model includes scarcity of resources, so that agents can explore the space just if they have enough resources to do it. Unlike the experiment, we assume variability of yields that follow over time and not a fixed pattern. Unlike the experiment, the space of possibility includes one million of securities of unknown yields. Yields are randomly distributed across space according to a function that is squeezed toward zero so that there are many low yield securities and fewer high yield securities. Unlike the experiment, the model is based on direct interaction among paired agents. Agents are randomly paired at the beginning of each simulation run. The number of interaction is 495, so that, on average, agents interact 5 times with each other.

	A1	A2	A3	Total
I1	12,50%	1,56%	12,50%	26,56%

I2	17,19%	4,69%	28,13%	50,00%
I3	1,56%	7,81%	14,06%	23,44%
Total	31,25%	14,06%	54,69%	100,00%

Table 15. Distribution of agents in the agent-based model according to experimental patterns.

In the model agents have an initial endowment of 1000 ECU and the exploration cost is fixed on 8000 ECU. This last has been kept as high as to reduce exploration frequency and to leave to agents enough time to build judgements on others, as though the scale time of cognitive evaluations was more frequent than that of space exploration. Table 16 summarises the main simulation parameters. Simulation runs are repeated 1000 times so that results are averaged.

Simulation Parameters	Values
Agents	100
Securities	1 million
Standard deviation of yields' distribution	500
Initial Endowment	1000 ECU
Exploration cost	8000 ECU
Number of interaction	495

Table 16. Simulation parameters.

The model we have built allows to point out how empirical foundations and analytical aims could be matched. In fact, we have based the largest part of agents' behaviour on data collected during the experiment, but we still miss some parts. Firstly, as mentioned before, the model consists of agents interacting each other in randomly assigned pairs that change every time steps, that is, on a very abstracted and implausible assumption. Secondly, we still do not have modelled how agents evaluate others' trustworthiness. The experimental evidence told us just that, on average, subjects were capable of identifying trustworthy partners, but we do not know how they did it. The analytic scenario introduced below can help us to investigate the systemic consequences of the theoretical hypotheses that we suggest to understand the possible mechanisms of trustworthiness.

Thus, we introduce seven simulation settings. In the first three we investigate general systemic outcomes and in the latter four we analyse some simple mechanisms for reputation formation. In the first setting (called "exploit_only") agents just exploit the securities randomly distributed at the beginning of the simulation. In the second one (called "explore_only") agents just randomly explore, when they have enough resources to do it. Such two sets are sort of baselines helping in the definition of the problem. In the third setting (called "listen_always") agents communicate information about their knowledge of the world as observed in the experiment, and they take decisions on what to do according to the experimentally based algorithms presented above. Partners are always considered as trustworthy. If agents want to explore new solutions by random or to exploit already known solutions, they do it, but if they want to listen to a suggestion about a new solution, they do it no matters whom is sending it. To say it differently, in "listen_always" agents always trust partners if they are saying something interesting.

In the following four settings, agents interact by forming a personal judgment on the reliability of the encountered agents as good/bad informers. Each agent is capable of recognising the agents she met before, to form an opinion on their reliability, and to remember it for future encounters. The differences between those four settings depend upon two issues: (i) how agents evaluate unknown partners or partners for whom the amount of positive feedbacks equals the one of negative ones; (ii) the possibility of sharing experience about partners reliability with others. The general rule for having feedbacks about partners trustworthiness is that when an agent meets a partner who showed to be unreliable, the agent records the cheater in her memory. Agents constantly up-to-date their memory according to a very simple rule: they compute the sum of true/false information that the agent encountered has transmitted in the past; if the sum of true information is higher than the total average of information transmitted, the agent is considered as reliable.

Thus, in the setting called "individual_J_pos", agents explore partners trustworthiness without sharing personal experience and with a "positive attitude" towards the other: when the paired agent has never met before, the counterpart is considered trustworthy and the information transmitted reliable until her action shows the contrary. On the contrary, in the setting called "individual_J_neg", agents consider unknown partners as unreliable ones, and the same happens when the number of positive past experiences with such a partner equals the number of negative ones. In sum, in this setting agents make use of personal past experience to form evaluations on reliability of agents so as to reduce the risk of being cheated. The last two sets, called "collective_J_pos" and "collective_J_neg", differ each other because of the different attitude towards unknown partners and work like the previous ones but introducing a memory at the system level that allows to share personal experiences: in this case, the image of agents (trustworthy/cheater) are made public, homogeneous, for all.

6. Simulation Results

Simulation results focus on the impact of the factors described above on main aggregate variables, such as the final profit of agents, their exploration capabilities and the stock of their resources. Figure 9 shows the best discovered security, that is, a proxy of the space exploration capabilities, and the endowment of agents in the baseline settings. As expected, while "exploit_only" implies incremental accumulation of resources but no capability of space exploration, "explore_only" implies no accumulation of resources, given the continuous space exploration followed by agents. The convex shape of the dynamics is due to the particular shape of the spatial distribution of securities mentioned above.

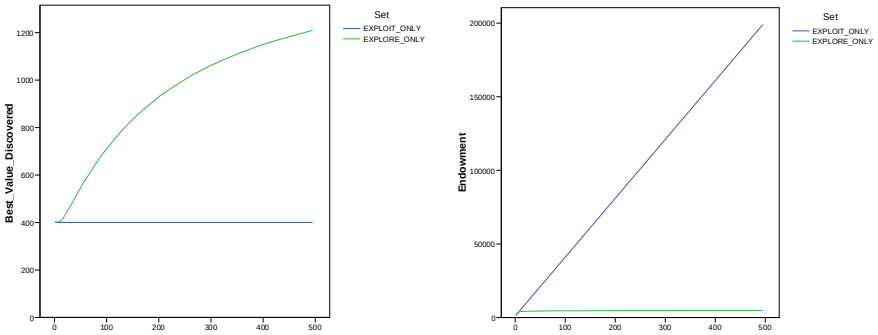


Figure 9. Dynamics of space exploration (on the left) and endowments of agents (on the right) in the baseline simulation settings.

Figure 10 points out the trade off considered by the uncertain system we are studying: exploration of the search space is expensive but necessary, and we are thus looking for mechanisms that allow efficiency on the side of costs and efficacy in finding better solutions (i.e, we are looking for higher levels of final profit which is a synthesis capable of measuring such a trade off).

Figure 14 shows that the set called "listen_always" is enough to reach levels of final profit higher than the "explore_only" case. To interpret such a result it is important to consider how the "listen_always" set is made. In fact, in such a simulation setting, the behaviour of agents is made of a mix of exploitation, random exploration and exploration driven by received information as we have observed in the experiment. But, here, we consider agents that are not capable of evaluating the trustworthiness of their partners, trusting all of them.

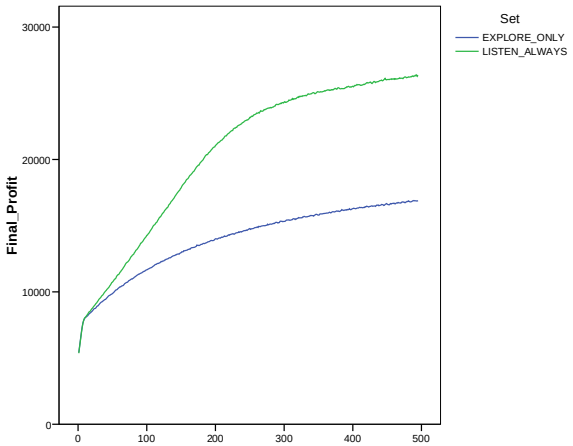


Figure 10. Dynamics of final profit of agents in "explore_only" and "listen_always" simulation settings.

The result shown in figure 11 is confirmed by the dynamics of space exploration and resources: the possibility to communicate, *per se*, means better systemic outcomes. This result is particularly noteworthy because it helps raise some methodological issues. As a matter of fact, in the "listen_always" setting, we have modelled action choices and decisions on the transmission of information to partners based upon experimental evidence by carefully conforming to experimental evidence. The same does not hold for the reception of the information, since we have explicitly ignored the robust experimental evidence about the fact that subjects clearly identified trustworthy subjects during the experiment. Moreover, it is highly plausible that the observed behaviour in experiments was strictly dependent upon the fact that subjects knew to be identifiable and were allowed for the evaluation of partners' trustworthiness. The consequence of this is that we have to go further and to consider settings such as "individual_J_pos", the results of which are depicted in figure 12.

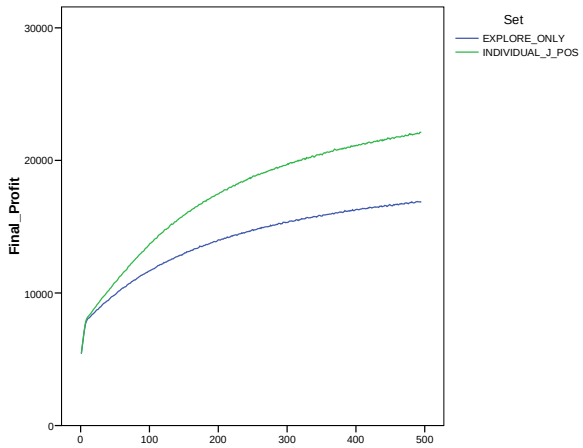


Figure 11. Dynamics of final profit of agents in "explore_only" and "individual_J_pos" simulation settings.

The figure 12 confirms that a reputation mechanism in which communication flows and is evaluated along the trustworthiness of partners is capable of guaranteeing better systemic outcomes. This is an interesting result, in particular given the fact that there is no general law or undisputed empirical evidence that suggests that reputation mechanisms allow people to cope with uncertain environments in a better way than other simplest micro mechanisms.

Another interesting evidence is that personal past-experience ("individual_J_pos") and shared past-experience ("collective_J_pos") generate about the same results, as shown in figure 16. The possibility to socially share the evaluations does not significantly increase the efficacy of the exploration capabilities at the macro level, nor it does the increased frequency of information availability. "Individual_J_pos" implies that each agent can form an evaluation of the other interacting agents just after completing the interaction with everybody. The process of evaluation formation is therefore slow and plodding. On the contrary, "collective_J_pos" implies that the information on agents is available to everybody since the end of the first interaction.

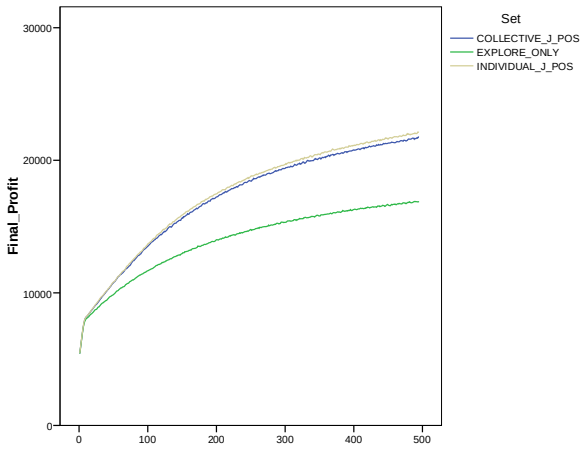


Figure 12. Dynamics of final profit of agents in "explore_only", "individual_J_pos" and "collective_J_pos" simulation settings.

Data on resources and on the exploration of search space that we do not graphically report here confirm that the positive result on the level of final profit is due to the capability of those two latter settings both in exploring the search space and in cumulating resources.

Furthermore it is worth verifying if these comparative results can be strongly influenced by the rule of "presumed innocence" that characterises the first dyadic interaction among strangers. Figure 13 shows the comparison of the values of final profits for the relevant simulation settings. The figure points out that a "negative" attitude towards the unknown further improves systemic outcomes[1]. The improvement allows for levels of final profit even higher than the hypothetical case of "listen_always" introduced before.

Figure 14 shows that such an improvement is almost all due to an increase in the accumulation of resources, while the exploration of space does not significantly differ depending upon such an attitude towards the unknown. The left part of figure 14 suggests a counterintuitive argument. Although the difference is not statistically significant, the positive attitude seems to guarantee a "better" exploration of the search space that is stable over time. Such a result contrasts with the idea that a negative approach towards unknown agents saves them in following wrong suggestions expressed by unreliable agents.

[1] The data of "collective_J_neg" simulation settings is not reported here but confirms preceding results: the sharing of past experiences does not have a significant impact on systemic outcomes and it does not change the improvement due to the adoption of a negative attitude towards unknown partners.

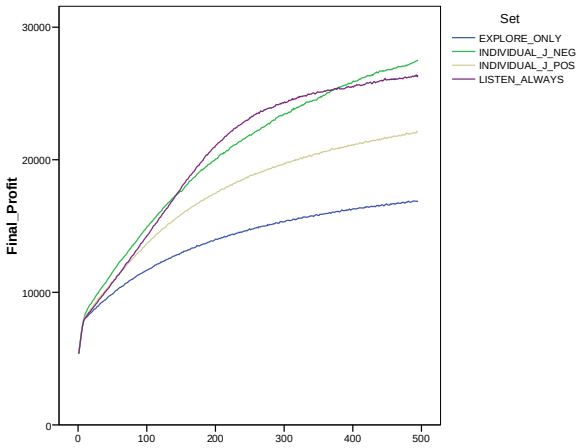


Figure 13. Dynamics of final profit of agents in "explore_only", "listen_always", "individual_J_pos" and "individual_J_neg" simulation settings.

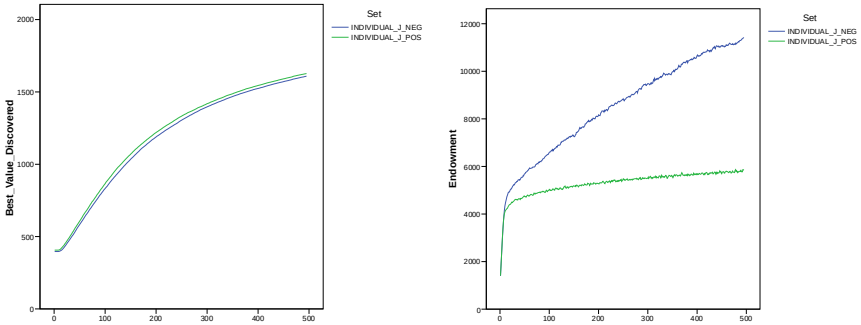


Figure 14. Dynamics of space exploration (on the left) and endowment of agents (on the right) in "individual_J_pos" and "individual_J_neg" simulation settings.

In order to check such an issue it is possible to refer to figure 15 where the average number of wrong suggestions that have been followed over time is represented (from now on we call them "lemons"). The figure shows that the average number of lemons in the system is very low, being less than 1 over 100 choices (in fact there are 100 agents choosing on each time step, but obviously not all the information flowing in the system is bad and it can become a "lemon"). The fact that the number of lemons decreases over time is, again, a not trivial outcome: Again, the point is a reputation mechanism is capable of decreasing the number of lemons in the system.

Another evidence is that the number of lemons in the case of the "individual_J_neg" setting is, in the first part of the simulation, higher than in the other case. This allows us to argue that the "negative" approach does not avoid lemons but the opposite, when agents have incomplete information about partners.

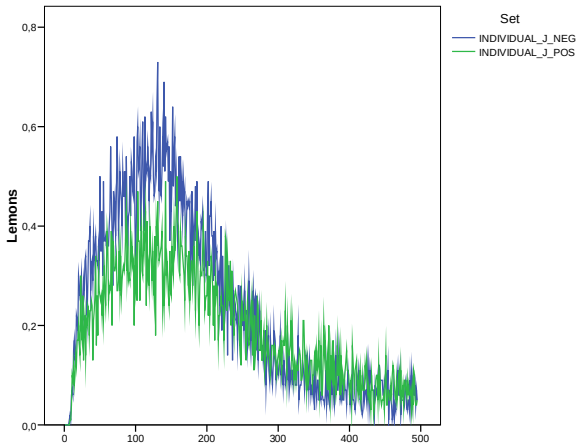


Figure 15. Dynamics of the average number of lemons in "individual_J_pos" and "individual_J_neg" simulation settings.

Figure 16 confirms the very slight effect of sharing experiences among agents: the number of lemons in the system is very similar to the case in which experiences are not shared. Furthermore figure 17 sheds light on some minor dynamic changes happening in the system only when a reputation mechanism is at work: in the case of the set "individual-J_neg" the number of true and false suggestions flowing in the system changes, showing a decrease in the number of suggestions in which a solution giving a very low yield is presented as very good (the case labelled "Higher_Low"). This is compensated in turn by an increase on the number of true suggestions about solutions that are good even if not the best ones found so far (i.e., the data series called "Other_Best").

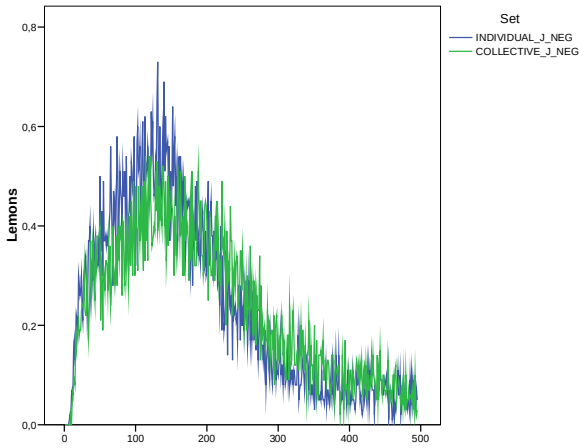


Figure 16. Dynamics of the average number of lemons in "individual_J_neg" and "collective_J_neg" simulation settings.

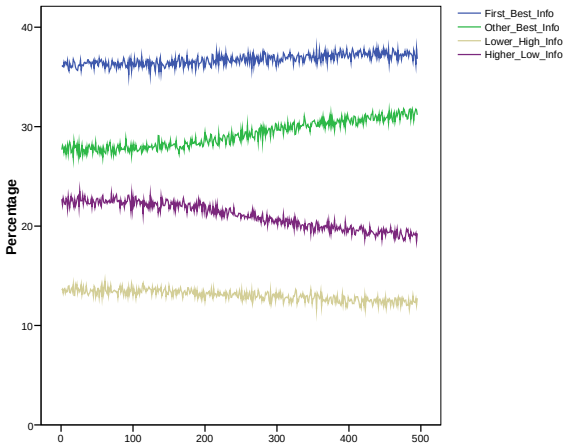


Figure 17. Dynamic distribution of kinds of information transmitted in "individual_J_neg" simulation settings.

Finally, a good question is whether there is a significant relationship between the economic performance of agents and the choice they made about the information they transmit to partners. Although the experimental data and the simulation exercise in introduced in par. 4 undoubtedly show that what agents decide to do with information does not directly depend upon their personal wealth, we reasonably expect that some indirect relationships mediated by systemic outcomes can exist. Thus, it is not surprising that, in results presented in figure 18, at the beginning of the simulation the average levels of final profit are similar for the different kinds of transmitted information. Just at the end of the considered period of time slight differences start to appear so that they can be interpreted as tendencies for the future: it seems that communicating the best solution is correlated with lower levels of final profit while the opposite is true when the cheating is at most (i.e., when a low yielding solution is presented as a very good one).

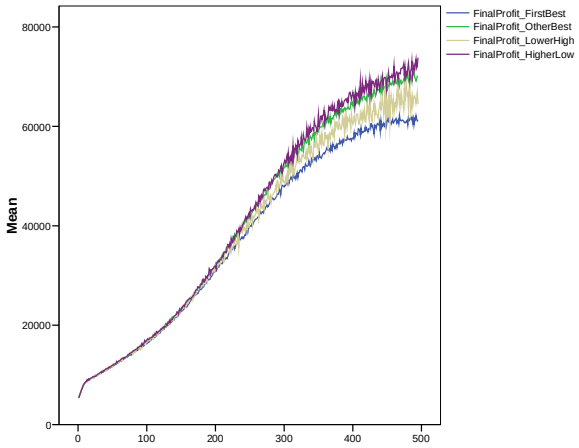


Figure 18. Dynamics of the average final profit of agents according to the kind of information they transmit in the "individual_J_neg" simulation settings.

In conclusion, it is worth re-stressing that these comparative results can be strongly influenced by the random dyadic interaction among agents. A first purpose for further simulations should thus be to explore interaction rules less rigid and exogenous, allowing agents to exploit trustworthiness as a means to choose partners, and to understand if they impact the resultant dynamics. A second development could be to introduce a further simulation setting where reputation of an agent is not made public for all but is conditioned by third-party mediated interaction. This would be a further step toward understanding the relevance of reputation mechanism in uncertain environments (Conte and Paolucci 2003).

7. Concluding Remarks

In recent years, reputation and trust have been subject of growing interest in many disciplines. Many computational models where reputation and trust have been modelled and theoretically investigated are available in the literature (e.g., Conte and Paolucci 2003; Lam and Leung 2006; Luke Teacy *et al*. 2006; Paolucci and Sabater 2006; Hahn *et al*. 2007). In this respect, the peculiarity of this paper is that we have introduced an experimental data-driven agent-based model where reputation is investigated starting from data on agents' behaviour generated through laboratory experiments (e.g., Boero and Squazzoni 2005; Janssen and Ostrom 2006). Rather than following a pre-constituted theoretical framework from which deriving assumptions, we have studied the impact of reputation on exploration capabilities of agents in uncertain environments by: i) observing the decision of human subjects in a laboratory experiment; ii) recognising patterns in experimental data; iii) validating the recognised patterns through a simulation that replicates the experiment; iv) reporting the validated patterns in a model and creating an experimental data-driven model to investigate relevant factors and mechanisms.

The first analytical result of our experimental driven simulation exercise is that reputation mechanism allows social agents to cope with an uncertain environment better than other 'pure' self-centred individual strategies. As said before, this is not a trivial result. In particular, as regards to the literature on exploration vs. exploitation in uncertain environments (e.g., March 1988, 1994; Holmqvist 2004; Sidhu, Volberda and Commandeur 2004), we have shown how much the inclusion of a reputation mechanism can enhance the exploration capabilities of agents. Of course, this is a *still in progress* work and future developments will be needed to elaborate more on this result, as said before. But, the results shown in this paper are based on experimental data and computationally robust enough to constitute some first sound evidences.

Acknowledgements

Financial support was provided by a FIRB 2003 grant (Strategic Program on Human, Economic and Social Sciences) from the Italian Ministry for the University and the Scientific Research [SOCRATE Research Project, coordinated by Rosaria Conte,

Protocol: RBNE03Y338_002]. A preliminary version of this paper has been presented at the Fifth European Social Simulation Conference, University of Brescia, 1-5 September 2008. We thank two ESSA 2008 reviewers and the audience for interesting intuitions and suggestions. We thank five anonymous conference referees for very helpful comments and remarks. Usual disclaimers apply.

References

Akerlof, G. A: The Market for Lemons: Quality Uncertainty and the Market Mechanism. Quarterly Journal of Economics, 84, 3, 488-500 (1970)

Beunza, D. and Stark, D.: The Organisation of Responsiveness: Innovation and Recovery in the Trading Rooms of Lower Manhattan. Socio-Economic Review, 1, 135-164 (2003)

Boero, R., Squazzoni, F.: Does Empirical Embeddedness Matter? Methodological Issues on Agent-Based Models for Analytical Social Science. Journal of Artificial Societies and Social Simulation, 8, 4: http://jasss.soc.surrey.ac.uk/8/4/6.html (2005)

Burt, R.: Brokerage and Closure: An Introduction to Social Capital. Oxford: Oxford University Press (2005)

Callon, M. (Ed.): The Laws of Markets. Oxford: Blackwell Publishers (1988)

Conte, R. and Paolucci, M.: Reputation in Artificial Societies: Social Beliefs for Social Order. Dordrecht: Kluwer Academic Publishers (2002)

Durlauf, S. N. and Young-Peyton H. (Eds.): Social Dynamics. Cambridge, MA: The MIT Press (2001)

Granovetter, M.: Getting A Job**:** A Study of Contacts and Careers. Harvard: Harvard University Press (1974)

Gui, B. and Sudgen, R. (Eds.): Economics and Social Interaction. Accounting for Interpersonal Relations. Cambridge: Cambridge University Press (2005)

Fischbacher, U.: z-Tree. Zurich Toolbox for Readymade Economic Experiments. University of Zurich, Working paper no. 21 (1995)

Hahn, C., Fley, B., Florian, M., Spresny, D., and Fischer, K.: Social Reputation: A Mechanism for Flexible Self-Regulation in Multiagent Systems. JASSS. 10(1): <http://jasss.soc.surrey.ac.uk/10/1/2.html> (2007)

Fischbacher, U.: z-Tree. Zurich Toolbox for Readymade Economic Experiments. University of Zurich, Working paper no. 21 (1995)

Holmqvist, M.: Experiential Learning Processes of Exploration and Exploitation Within and Between Organizations: An Empirical Study of Product Development. Organization Science, 15, 1, 70-81 (2004)

Janssen, M. and Ostrom, E.: Empirically Based, Agent-Based Models. Ecology and Society, 24, 33-60 (2006)

Knorr Cetina, K. D. and Preda, A. (Eds.): The Sociology of Financial Markets. Oxford: Oxford University Press (2004)

Lam, K. M. and Leung, H. F.: A Trust/Honesty Model with Adaptive Strategy for Multiagent Semi-Competitive Environments. Autonomous Agents and Multi-Agent Systems, 12, 293-359 (2006)

Luke Teacy, W. T., Patel, J, Jennings, N. R. and Luck M.: TRAVOS: Trust and reputation in the context of inaccurate information sources. Autonomous Agents and Multi-Agent Systems, 12, 293-359 (2006)

March, J.: Decisions and Organizations. Oxford: Basil Blackwell (1988)

March, J.: A Primer on Decision Making: How Decisions Happen. New York: The Free Press (1994)

Paolucci, M. and Sabater, J.: Introduction to the Special Issues on Reputation in Agent Societies. Journal of Artificial Societies and Social Simulation, 9, 1: <http://jasss.soc.surrey.ac.uk/9/1/16.html> (2006)

Raub, W. and Weesie, J.: Reputation and Efficiency in Social Interactions: An Example of Network Effects. The American Journal of Sociology, 96, 3, 626-654 (1990)

Sabater, J. and Sierra, C.: Review on Computational Trust and Reputation Models. Artificial Intelligence Review, 24, 33-60 (2005)

Sidhu, J. S., Volberda, H. V., and Commandeur, H. R.: Exploring Exploration Orientations and Its Determinants: Some Empirical Evidence. Journal of Management Studies, 41, 6, 913-932 (2004)

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., and Milinski, M.: Gossip as an alternative for direct observation in games of indirect reciprocity. PNAS 104(44), 17435-17440 (2007)

Williamson, O. E.: Transaction Cost Economics: The Governance of Contractual Relations. Journal of Law and Economics, 22, 233-261 (1979)

Embedded Trust: An Experiment on Learning and Control Effects

Vincent Buskens, Werner Raub, and Joris van der Veer

Version December 10, 2008; file: emb_trust_exp_v101208

Department of Sociology / ICS, Utrecht University
Heidelberglaan 2, 3584 CS Utrecht, Netherlands
v.buskens@uu.nl, w.raub@uu.nl

Abstract
The paper discusses a laboratory experiment in which pairs of trustors play finitely repeated Trust Games with the same trustee. We study trustfulness of the trustors as well as trustworthiness of the trustee. We distinguish between learning and control effects on trustfulness and trustworthiness. Learning effects are related to an actor's information on past behavior of the partner. Control effects are related to opportunities for sanctioning a trustee in future interactions. The experiment includes two conditions that represent different types of "embeddedness" of Trust Games. In one condition, each trustor only knows what happens in her own games with the trustee. In the other condition, each trustor also knows what happens in the games of another trustor with the trustee. Thus, with respect to trustfulness of the trustor, the design allows for disentangling learning effects from own experience of the trustor with the trustee and learning effects through third-party information, i.e., information on experiences of the other trustor with the trustee. Also, the design allows for disentangling control effects on trustfulness and on trustworthiness through own sanction opportunities of the trustor and through opportunities for third-party sanctions, i.e., sanctions implemented by the other trustor.

1 Introduction

Social and economic exchange often presupposes trust between actors. When lending a book to a colleague, we trust the colleague to return the book in good shape. A buyer of a second-hand car trusts the dealer to be honest about hidden defects of the car. We may be more inclined to trust a colleague who has often returned our books in good shape in the past. The more information you have from your friends on their good experiences with a second-hand car dealer, the more you may be inclined to buy a car yourself from the dealer unless, maybe, you happen to know that the dealer is about to retire and close down his outlet. These examples illustrate the intuition that trustfulness may be fostered by positive information about trustworthiness of the trustee in the past. The last example illustrates that trustfulness and trustworthiness might become problematic when opportunities for future sanctioning of the trustee's present behavior become infeasible.

Our examples are related to the "embeddedness" of trust problems and exchange in the sense of [17]. Embeddedness refers to repeated transactions over time between the same partners and to transactions between partners who share a network with third parties. In [7], Buskens and Raub distinguish two mechanisms through which embeddedness affects trust[1]: learning and control. On the one hand, actors can learn that a trustee has been trustworthy in the past and may infer from this that the trustee is likely to be trustworthy now as well. On the other hand, actors can base trust on sanction opportunities in the future. The more extensive future sanctions can be, the more likely that the trustee realizes that his short-term incentives from abusing trust do not compensate for the long-term losses he will incur due to the sanctions of the trustor. Consequently, the more extensive the sanction opportunities of the trustor, the more likely it is that the trustor can trust the trustee because it is more likely that the trustee will be trustworthy. Buskens and Raub also distinguish between two levels of embeddedness: the level of the dyad and the network level. They argue that learning and control operate at both of these levels (see also [27, pp 138-139] for a similar discussion of learning and control through network embeddedness). Trustors can learn through own experiences and through experiences of others. Sanctions can be executed by the trustor herself or by third parties such as other trustors of the trustee.

Learning and control effects on trust through dyadic and network embeddedness are intimately connected to reputation effects on trust. Roughly (see, e.g., [22, pp. 629-633] for an extensive discussion), the reputation of an actor is a characteristic or an attribute that partners ascribe to the actor. The empirical basis of an actor's reputation is his observed past behavior. Thus, the trustee's present reputation for trustworthiness ascribed to him by a given trustor depends on the trustor's own prior experiences with the trustee as well as on third-party information on the trustee that the trustor receives from other trustors. The trustee's present reputation for trustworthiness thus depends on dyadic learning as well as on network learning of the trustor and will affect the trustor's trustfulness. In addition, the trustee's present trustworthiness will affect his future reputation for trustworthiness vis-à-vis the trustor with whom the trustee interacts today as well as vis-à-vis other trustors who receive information on the trustee's present behavior. Reputation effects on trust can thus be conceived as learning and control effects on trust through dyadic and network embeddedness.

Experimental as well as survey research provides evidence for effects of embeddedness on trust (see [8] for an overview). The problem with most of the evidence is that effects of learning and control are hard, if at all, to disentangle. Therefore, *how* embeddedness affects trust is a largely unresolved question. This paper provides new evidence on learning and control effects on trust through dyadic embeddedness and network embeddedness from a laboratory experiment in which subjects have to make incentive-guided choices. The experiment complements earlier research in four ways. First, learning and control mechanisms can be clearly disentangled. Second, the design of the experiment ensures that embeddedness characteristics are exogenously given rather than being themselves results of individual choices. This facilitates the interpretation of empirical findings on embeddedness effects. Third, subjects are pre-

[1] The noun "trust" is used as shorthand for "trustfulness *and* trustworthiness." Only when used as a verb or in conjugations such as "abuse trust" and "honor trust," trust refers exclusively to "trustfulness."

cisely and truthfully informed on the incentive structure and their payments depend on their own and others' choices. While the external validity of experimental designs remains questionable in principle, the assumption of external validity becomes plausible to a considerable extent when results that have been found employing other research designs such as surveys are replicated in this experiment. Finally, we analyze effects of embeddedness through learning and control not only on trustfulness of the trustor but also on trustworthiness of the trustee.

2 Embedded Trust: Hypotheses

We define a trust situation as an interaction with strategic interdependence between two actors, the trustor and the trustee. If the trustor trusts the trustee, the trustee has the possibility and an incentive to act opportunistically, i.e., to abuse trust. Compared to the situation when she does not trust the trustee, the trustor regrets being trustful if trust is abused but she is better off after trusting the trustee if the trustee does not behave opportunistically. A trust situation can be represented as the well-known Trust Game shown in Figure 1. The Trust Game is a game-theoretic representation of a trust situation [11, 12, 13, 20]. The Trust Game starts with a move by the trustor, who trust (i.e., she is "trustful") or does not trust. If the trustor does not trust, the game is over, with trustor and trustee each obtaining a payoff P. If the trustor trusts, the trustee chooses between honoring trust (i.e., he is "trustworthy") and abusing trust. If the trustee honors trust, trustor and trustee each receive $R > P$. If the trustee abuses trust, the trustor receives $S < P$, while the trustee receives $T > R$.

In our experiment, Trust Games are played in triads comprising two trustors and a trustee. Clearly, a triad represents a small network between the actors. First, one of the trustors, say, trustor 1, plays a Trust Game with the trustee. After this Trust Game has been finished, the other trustor, trustor 2, plays a Trust Game with the same trustee. This pair of two games is played 15 times. All actors, trustors and trustee, have complete information about the whole structure of the game such as the number of

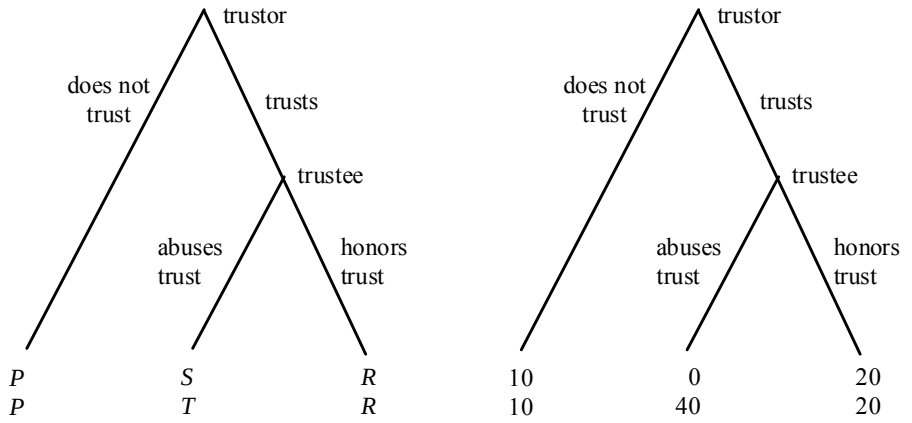


Fig. 1. Extensive form of a Trust Game. $T > R > P > S$. The right-hand Trust Game is the numerical example used in the experiment

rounds to be played, each actor's payoff function, etc. The experiment employs two information conditions and actors know in which information condition they are:

1. *No information exchange between trustors*: each trustor only knows what happens in her own Trust Games with the trustee but is not informed about what happens in the games of the other trustor playing with the same trustee.
2. *Full information exchange between trustors*: after each Trust Game, also the trustor not involved in that game receives information on the choices made in that game.

Assume now that subjects differ in the way they value outcomes because of their social orientations [24, 26] or preferences for fairness or reciprocity [3, 14, 21]. This means that actors' utilities may differ from their own material payoffs. Also, while knowing their own utility function, actors may be incompletely informed on the utility functions of other actors. Thus, our basis for learning is the trustors' uncertainty about trustees' utility from the payoffs in the games and the possibility that these utilities might be such that some trustees do not have an incentive to abuse trust. A detailed discussion of theoretical models that include such assumptions is beyond the scope of this paper (see, e.g., [6, 11, 13]). However, it is intuitively clear that trustors might now believe that some trustees will not abuse trust. If trustors have doubts about the behavior of at least some trustees, this can have consequences even for those trustees who do have an incentive to abuse trust. Namely, such trustees can profit from appearing trustworthy in early rounds of a finitely repeated Trust Game and can thus have an incentive for reputation building, while they will abuse trust towards the end of the game.

We can now derive hypotheses on learning effects through embeddedness on trustfulness of trustors from various backward looking learning models (see [10, 16] for overviews of such models). They typically imply (see [9] for details) that trustfulness is more likely if trust has been honored more frequently and also, accounting for discounting of past experiences, more recently. On the other hand, trustfulness will become less likely after trust has been abused. The behavior of the trustee is expected to be largely determined by his concern to build and keep up a good reputation and, therefore, is expected to be mainly driven by control effects. We thus test the following two hypotheses on learning effects on trustor behavior:

Hypothesis 1 (dyadic learning): The more a trustor's trustfulness has been honored in the past, the more likely it is that this trustor is trustful; the more a trustor's trustfulness has been abused in the past, the less likely it is that this trustor is trustful.

Hypothesis 2 (network learning): The more information a trustor has that trustfulness of another trustor has been honored in the past, the more likely it is that she will be trustful herself; the more information a trustor has that trustfulness of another trustor has been abused in the past, the less likely it is that she will be trustful herself.

In addition to learning effects and in line with arguments on incentives for reputation building for the trustee, we can derive hypotheses on control effects. The theoretical basis for these hypotheses are models for games with incomplete information [19] that assume rational forward looking behavior with learning in the sense of Bayesian

updating of beliefs (see [4, 6, 11, 13] for applications of such models to finitely repeated Trust Games). The basic intuition is that a trustee will be more likely to be trustworthy and, therefore, the trustor more likely to be trustful, the more the trustee has to lose in future games after he would abuse trust. The losses come from trustors not being trustful anymore because they experienced abused trust themselves or hear about it from another trustor. Trustor 2 can profit less from this network information at the end of the game because there is one game less to be played after her game compared to the trustor 1's game in the same round. This leads to the following hypotheses on control effects on trust and trustworthiness (see [9] for details):

Hypothesis 3 (dyadic control – trustor): The more rounds left in the game, the higher the likelihood that a trustor is trustful; the likelihood of trustfulness decreases faster in the last few rounds of the game than in earlier rounds (end-game effect).

Hypothesis 4 (network control – trustor): In the condition with full information exchange between trustors, compared to the condition with no information exchange, the likelihood of trustfulness is higher and will decrease less in early rounds. The end-game effects will be stronger for trustor 2 than for trustor 1.

Hypothesis 5 (dyadic control – trustee): The more rounds left in the game, the higher the likelihood that trust will be honored; the likelihood of trustworthiness decreases faster in the last few rounds of the game than in earlier rounds (end-game effect).

Hypothesis 6 (network control – trustee): In the condition with full information exchange between trustors, compared to the condition with no information exchange, the likelihood of trustworthiness is higher and will decrease less in early rounds. The end-game effect will be stronger in games with trustor 2 than in games with trustor 1.

3 The Experiment

In the experiment, the outcomes of the Trust Games are points that subjects earn. If the trustor is not trustful, this yields 10 points for both trustor and trustee; when trust is honored, each actor receives 20 points; when trust is abused, the trustee receives 40 points, leaving the trustor with no points (see the right-hand Trust Game in Figure 1). Subjects are paid 1 eurocent for each point they earn at the end of the experiment.

The experiment was programmed using z-Tree software [15]. Subjects play the Trust Game in supergames of 15 rounds. Subjects are matched in groups of three, one trustee and two trustors, which we call triads. In each of the 15 rounds, the trustee plays one Trust Game with each of the two trustors. During the 15 rounds, the trustee plays with the same two trustors in each round. Therefore, in every round, while trustors play one Trust Game, the trustee plays two Trust Games, adding to 30 Trust Games played per supergame by one trustee. The trustee not necessarily needs to make a choice in all 30 games: when the trustor does not trust, the trustee has no choice to make.

In every round, the trustee always plays with the same trustor first, while the other trustor has to wait, and always plays second with this trustee. Thus, within a supergame, the trustors always move in the same order. The trustors are referred to as trustor 1 and trustor 2, respectively.

As already mentioned, there are two conditions in the experiment. In both conditions, the trustee is immediately informed on the trustor's move in the current Trust Game. Between conditions the amount of information is varied that is shared among the two trustors playing with the same trustee. In the "no information exchange between trustors" condition, trustors do not share any information. In the "full information exchange between trustors" condition, trustors playing with the same trustee do share all information about each other's games. In this condition, as soon as a game has been played, either the first or the second game in the round, also the other trustor receives information on the choices made in this game. Information is provided automatically by the computer and is always truthful. All subjects know in what information condition they are and thus also know what information is available to the other two actors in their triad. All subjects see the outcomes of the games they played themselves in previous rounds on their screens. In the condition with full information exchange, each trustor also sees the outcomes of the previous Trust Games of the other trustor on the screen. Note that the experimental design ensures that embeddedness is exogenous and is not itself a result of subjects' choices.

Every subject played three times a supergame as described above, once as a trustee, once as trustor 1, and once as trustor 2. Each subject played all three supergames in the same information condition. In between the three supergames, the subjects were rematched to other subjects. Subjects were never rematched to other subjects they already played with in a previous supergame. This was made common knowledge to all subjects.

The experiment was conducted in the ELSE laboratory of Utrecht University, in a computer room specially designed for experiments: every subject was seated behind a PC in his or her own cubicle, with a separate cubicle for the experimenters. Each session comprised 18 subjects. When all 18 subjects were present, instructions on paper were provided and the treatment was started. Within a session, the instructions were the same for everyone and this was also made common knowledge to the subjects. Subjects were explained how the game worked they were about to play, and that they would receive 1 eurocent for every point they earned.

After reading the instructions, a few questions were asked on the computer screen so that subjects could check whether they understood the instructions. In case of incorrect answers, subjects were provided the correct answers and a brief explanation. Subjects also played two practice rounds in which they could not earn any points yet. These rounds were not played with other subjects, but with the computer to guarantee standardized moves of the "partners" in these rounds.

Then the three supergames were played. At the end of the experiment, subjects had to fill in a short questionnaire, including items concerning their general trustfulness and trustworthiness. Meanwhile, the experimenters prepared closed envelopes with the earnings for each subject. The sessions took between 55 and 70 minutes. Subjects earned on average €10.67, with an average of €11.10 for subjects in the condition with full information exchange and €10.25 for those in the no information exchange condition. The minimum and maximum earnings were €7 and €12.40.

The information about the structure of the experiment such as the number of rounds, roles (i.e., being trustor or trustee), and what subjects would get to know was honestly provided and subjects were never deceived or in another situation than told. In order to prevent inducing normative associations, the names of the different roles and their possible moves were rendered neutrally. For instance, the moves of the trustee were labeled "down" and "right" rather than "honor trust" and "abuse trust."

In total, 72 subjects participated in the experiment, 28 male and 44 female, mostly undergraduate students from different fields, most of them students of social sciences. Subjects were recruited using the online recruitment system ORSEE [18]. Four sessions were scheduled and 18 subjects participated in each session. Two sessions were played in the condition with no information exchange between trustors and two sessions in the condition with full information exchange.

4 Data and Statistical Model

The experiment comprised four sessions, with six triads per session and three supergames of fifteen rounds per subject, each round comprising two Trust Games. Thus, $4\times6\times3\times15\times2 = 2160$ Trust Games were played in total. Note that trustee behavior is observed only in those games in which the trustor is not trustful. There are 485 games in which there was no trust, leaving 1675 games (78% of the total number of games played) in which the trustee's behavior is observed. Trustors are more trustful in the full information exchange condition (in 913 of the 1080 games) than in the no information exchange condition (in 762 of the 1080 games) and trustfulness decreases over rounds. Also, trustworthiness decreases over the rounds. These descriptive findings are in line with our hypotheses. Figure 2 displays the descriptives graphically.

The descriptives are in line with earlier experiments (see [8, 10, pp. 446-453] for overviews). Trustfulness and trustworthiness are high for most of the rounds, with strong end-game effects in the last couple of rounds. We observe that repeating the Trust Game with the same partner, i.e., dyadic embeddedness (also known as "partner matching" in the experimental literature), and the availability of third-party information, i.e., network embeddedness, have complementary effects. This could not be concluded from otherwise closely related earlier experiments [2]. Summarizing, the descriptive analyses show that dyadic embeddedness leads to rather high levels of trustfulness and trustworthiness, higher than is normally found in one-shot Trust Games, including series of one-shot Trust Games, each played with a different partner (also known as "stranger matching" in the experimental literature). Still, additionally providing information about Trust Games of the trustee with another trustor induces even more trustfulness and trustworthiness.

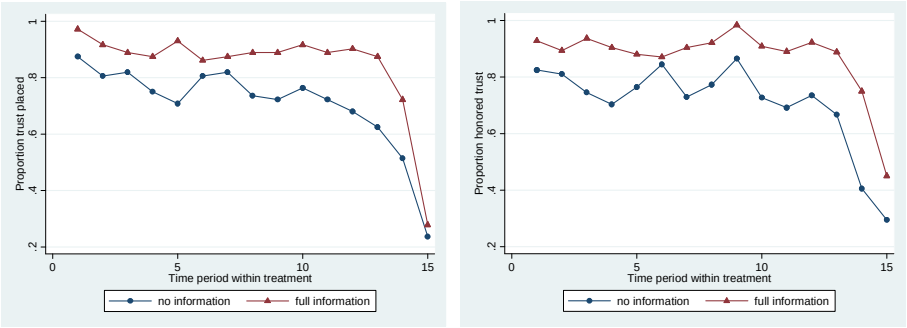


Fig. 2. Proportion of games in which trust was trustful (left) and trust was honored (right), per round, and per experimental condition.

Our dependent variables identify the behavior of the trustor and the trustee. Various independent variables represent past experiences. First, for each round t, we construct the weighted number of times a trustee honored trust in the past as

$$\text{PASTHONOR\_OWN} = \sum_{\tau=1}^{t-1} w_1^{\tau}\,\text{HONOR\_OWN}_{\tau}. \tag{1}$$

We define PASTHONOR_OWN = 0 for the first round of a supergame. Similarly, we define a variable for the number of times trust was abused in the past:

$$\text{PASTABUSE\_OWN} = \sum_{\tau=1}^{t-1} w_1^{\tau}\,\text{ABUSE\_OWN}_{\tau}. \tag{2}$$

Note that ABUSE_OWN and HONOR_OWN can be equal to 0 simultaneously, namely, when the trustor is not trustful. Therefore, the effects of the two variables just defined should be interpreted relative to the number of times the trustor was not trustful. The parameter w_1 is estimated in the statistical model. We assume that w_1 is the same for honored trust and for abused trust in the past.

In the condition with full information exchange, each trustor also receives information about the games of the other trustor. To complete the set of independent variables needed for testing hypotheses on learning effects for the trustor, we thus define variables for such third-party information:

$$\text{PASTHONOR\_OTHER} = \sum_{\tau=1}^{t-1} w_2^{\tau}\,\text{HONOR\_OTHER}_{\tau};\ \text{PASTABUSE\_OTHER} = \sum_{\tau=1}^{t-1} w_2^{\tau}\,\text{ABUSE\_OTHER}_{\tau}. \tag{3}$$

Information from the other trustor might be forgotten or discarded faster than own experiences. Therefore, we introduce the parameter w_2. The variables PASTHONOR_OTHER and PASTABUSE_OTHER are always set to 0 in the condition without information exchange between trustors.

The variables representing control effects are straightforward. FUTURE is the number of rounds left. FUTUREFULL represents network control. This variable is the inter-

action of FUTURE with a dummy for whether subjects are in the condition with full information exchange between trustors. In addition, we use dummies that indicate the last but one and the last round of the repeated game: ROUND14, ROUND15. These variables are again interacted with dummies for the information condition: ROUND14FULL, ROUND15FULL. We also distinguish between the games with trustor 1 and with trustor 2 in the last two rounds for the information condition. Therefore, two dummies are used for the games with trustor 2 in the full information condition in these two rounds: ROUND14TR2FULL, ROUND15TR2FULL.

The statistical model used to analyze the data is a three-level logistic regression model. Based on the difference in attractiveness between the two possible moves of the trustor and assuming that there is some randomness in the extent that we know the attractiveness, we can estimate a logistic regression model for the probability that trustor *j* in triad *i* is trustful at time *t*:

$$P_{ijt}^{Trustful} = \frac{e^{\beta_1\cdot(\text{own learning})+\beta_2\cdot(\text{TP-learning})+\beta_3(\text{own control})+\beta_4(\text{TP-control})+u_i+v_{ij}+\varepsilon_{ijt}}}{1+e^{\beta_1\cdot(\text{own learning})+\beta_2\cdot(\text{TP-learning})+\beta_3(\text{own control})+\beta_4(\text{TP-control})+u_i+v_{ij}+\varepsilon_{ijt}}}, \quad (4)$$

where the β's indicate the vectors of regression coefficients for the respective groups of independent variables, namely, variables representing own learning, learning through third-party information, own control opportunities, and control opportunities involving third parties. Furthermore, u_i is a random component for the triad in which the decision is made, v_{ij} is a random component for the trustor within the triad who makes the decision, and ε_{ijt} is the random component for each individual decision. This is a hierarchical three-level model (see, e.g., [25]). Strictly speaking, we have a cross-classified nesting because trustors are involved twice in a series of 15 Trust Games with different partners. However, estimating this more complex structure affects the outcomes of the analyses only marginally. Also, the random component related to trustors has a similar size in this more complex estimation.

The statistical model for the probability that trustee *i* in his triad honors trust with trustor *j* in this triad at time *t* looks rather similar:

$$P_{ijt}^{Honor} = \frac{e^{\beta_1(\text{control})+\beta_2(\text{TP-control})+u_i+v_{ij}+\varepsilon_{ijt}}}{1+e^{\beta_1(\text{control})+\beta_2(\text{TP-control})+u_i+v_{ij}+\varepsilon_{ijt}}}, \quad (5)$$

where the β's again represent the regression coefficients; u_i is a random component for the trustee making the decision. In this case, this is equivalent to the triad in which the decision is made; v_{ij} is a random component for the trustor in the triad with whom the trustee is playing a specific Trust Game; and ε_{ijt} is the random component for each individual decision. Again, the specification of the random components could have been more complex, because each trustor is involved in two triads and, therefore, the random component for trustors could have been specified as a cross-classified model in which the random component represents randomness related to a specific subject playing as a trustor. Because random components related to the trustors are consistently estimated to be 0 in the models for explaining trustees' behavior, we stick to the simpler hierarchical model in which we control for nesting of trustees' decisions within trustors.

5 Results

We first analyze trustfulness of trustors. As a baseline model, we report a logistic regression of the likelihood that a trustor is trustful with a dummy (FULL INFORMATION) for the experimental condition, a dummy for whether trustor 2 is involved rather than trustor 1 (TRUSTOR2), and two dummies for the second and the third supergame (TREATMENT2, TREATMENT3) in Table 1. This model shows that there is more trustfulness in the condition with full information exchange between trustors than in the condition with no information exchange. Trustfulness also increases over the supergames, as indicated by the significant difference between the first and the third supergame, while the second supergame is in between. The random parts at the triad and trustor level show that about 28% of the variance can be attributed to the triad, while

Table 1. Three-level logistic regression of the likelihood for a trustor to be trustful (2160 decisions by 144 trustors in 72 triads).

		Baseline model		Full model	
	Hyp.	Coeff.	Std. err.	Coeff.	Std. err.
FULL INFORMATION	+	0.983**	0.332	-0.691	0.625
TREATMENT2		0.381	0.379	-0.031	0.251
TREATMENT3		0.911*	0.385	0.448	0.282
TRUSTOR2		-0.059	0.124	-0.031	0.192
PASTHONOR_OWN	+			1.788**	0.176
PASTABUSE_OWN	−			-1.539**	0.238
PASTHONOR_OTHER	+			0.746**	0.250
PASTABUSE_OTHER	−			-1.142**	0.396
FUTURE	+			0.076**	0.029
FUTUREFULL	−			0.059	0.050
ROUND14	−			-0.791*	0.385
ROUND15	−			-1.943**	0.439
ROUND14FULL				-0.424	0.778
ROUND15FULL				-1.742*	0.738
ROUND14TR2FULL	−			-0.284	0.825
ROUND15TR2FULL	−			0.274	0.728
ROUND1TREATMENT1				1.369*	0.559
ROUND1TREATMENT2				1.621**	0.615
ROUND1TREATMENT3				1.838*	0.810
CONSTANT		0.784*	0.321	-0.207	0.386
Variance triad level		1.364	0.332	0.150	0.154
Variance trustor level		0.251	0.114	0.304	0.194
Variance decision level		3.290		3.290	
Loglikelihood		-972.51		-616.38	

*, ** indicate significance at $p < 0.05$ and $p < 0.01$, respectively (two-sided tests).

only 5% can be attributed to a specific trustor within a triad. Through separate analyses we found that almost all the variance at the trustor level is due to the no information exchange condition, while there is hardly any variance that can be attributed to the individual trustors in the full information exchange condition.

In the full model, the main effect of the full information condition vanishes, indicating that the difference between the experimental conditions is mainly due to the learning and/or control variables in the full model. The full model provides clear evidence for Hypothesis 1 on effects of dyadic learning as well as for Hypothesis 2 on network learning. Trustors are more trustful after experiencing themselves more honored trust and they are less trustful after experiencing themselves more abused trust. In addition, when they observe that the other trustor's trustfulness is honored more often, this also increases their likelihood to be trustful themselves, while their own trustfulness decreases if they observe more abused trust of the other trustor in the same triad. In addition, it can be seen that the effect of experiencing honored trust in the trustor's own games is larger than the effect of information about honored trust in games of the trustee with the other trustor. It is striking that the effect of information about abused trust in games of the trustee with the other trustor is almost as large as the effect of experiencing abused trust in the trustor's own games with the trustee. The significance of the effect of information about abused trust in games with the other trustor is smaller as can be inferred from the larger standard error, but that could be due to the fact that there are less data on these experiences because they only occur in the condition with full information exchange.

Considering control effects on trustfulness, we see a clear dyadic control effect of the rounds still to be played. Also, the end-game effects are strong, starting in round 14, while dummies for earlier rounds did not add to the explained variance. These results provide support for Hypothesis 3. However, there is not much evidence for Hypothesis 4 about of network control effects on trustfulness. The interaction of the number of rounds left with the full information exchange condition is not significant, indicating that the general decrease of trust is not less strong in the condition with full information exchange between trustors. In addition, there is no main effect left of the full information condition after controlling for learning, which would indicate a network control effect. The end-game effects are about twice as strong in the full information condition as in the condition with no information condition and this difference is significant for round 15. These steeper network effects are mainly due to earlier experiences of honored trust in the condition with full information exchange through which the level of trustfulness is higher before it starts to decrease. The two additional dummies that interact the variables indicating the two final rounds in the full information condition with a variable indicating the trustor who is playing are not significant. This implies that we also do not find evidence for the second part of Hypothesis 4. As we will see below, the trustee does anticipate on the network control opportunities of the trustors, which is also the main explanation why trust remains higher in the full information condition, but the trustors do not seem to anticipate themselves on this anticipation of the trustee. Finally, the controls for the first rounds show that the starting level of trust slowly increases over the treatments, indicating that trustors realize more and more that trustfulness can be beneficial in the beginning of a supergame.

We now turn to the analysis of the trustworthiness of trustees. The baseline model in Table 2 shows that there is more trustworthiness in the condition with full informa-

tion exchange between trustors than in the condition without information exchange. While trustors develop some more trustfulness over the treatments, it is not the case that the trustees' behavior changes significantly over the treatments. Controlling for these treatment variations, we see that no unexplained variance is attributed to the trustor level and that 28% of the unexplained variance is attributed to the trustee level. The remaining 72% of the unexplained variance is at the decision level.

The full model provides the results of the tests of the hypotheses on trustworthiness. Notice that the difference between the two experimental conditions is not explained away by the hypothesized effects for the trustee. With respect to control effects on the likelihood of trustworthiness, the effect of dyadic embeddedness (FUTURE) is strongly significant. However, the interaction term with the full information condition (though non-significant) shows that dyadic control is only present in the condition without information exchange between trustors. Apparently, control is so strong in the condition with information exchange (which is also indicated by the remaining main effect of full information) that the likelihood of trustworthiness remains at or even above the 90%-level throughout the first thirteen rounds of a supergame. The fact that the main effect of full information is positive and that trustworthiness does not decrease in the first 13 rounds in the full information condition thus indicates that there is an additional control effect of network embeddedness over and above dyadic control. After round 13, there is a clear drop in the likelihood of trustworthiness in both information conditions. When we study these end-game effects in

Table 2. Three-level logistic regression of the likelihood to honor trust (1542 decisions with 144 trustors by 72 trustees in 72 triads).

		Baseline model		Full model	
	Hyp.	Coeff.	Std. err.	Coeff.	Std. err.
FULL INFORMATION	+	1.405**	0.317	2.263**	0.576
TREATMENT2		0.295	0.385	0.314	0.510
TREATMENT3		0.439	0.383	0.378	0.507
TRUSTOR2		-0.051	0.147	0.004	0.164
FUTURE	+			0.080*	0.031
FUTUREFULL	−			-0.072	0.050
ROUND14	−			-1.914**	0.463
ROUND15	−			-2.588**	0.704
ROUND14FULL				1.292	0.823
ROUND15FULL				-0.371	1.095
ROUND14TR2FULL	−			-1.798*	0.810
ROUND15TR2FULL	−			-2.764*	1.217
CONSTANT		0.681*	0.318	0.305	0.480
Variance triad level		1.289	0.352	2.474	0.658
Variance trustor level		0	0	0	0
Variance decision level		3.290		3.290	
Loglikelihood		-657.32		-593.65	

*, ** indicate significance at $p < 0.05$ and $p < 0.01$, respectively (two-sided tests).

more detail, we see that the trustee is much less trustworthy with trustor 2 in the full information condition than with trustor 1. This indicates an additional network control effect, because trustor 2, compared to trustor 1, has less (or no) control opportunities especially in these last two rounds. Summarizing, these results provide quite some support for Hypothesis 5 as well as for Hypothesis 6.

6 Conclusion and Discussion

In this paper, we have discussed an experiment in which pairs of trustors play Trust Games with the same trustee. This is the simplest set-up for simultaneously studying effects of dyadic embeddedness and network embeddedness on trust. We distinguished between learning and control effects of embeddedness. We have analyzed how both trustfulness and trustworthiness are affected by embeddedness.

Learning effects at the dyadic and the network level are both strong determinants of trustfulness of trustors. We also find dyadic control effects on trustfulness, but we do not find evidence for network control effects on trustfulness. The higher levels of trustfulness under network embeddedness are actually caused by the trustees anticipating on the stronger sanction opportunities of the trustors. The trustees are more trustworthy under network embeddedness, which has the consequence that the trustors have more positive learning experiences leading to more trustfulness. The effects on trustfulness are very consistent with earlier findings [7, 8]. For trustees, we find dyadic control effects as well as network control effects on trustworthiness.

We conclude by briefly returning to our findings that network control opportunities affect trustee behavior while there is evidence for dyadic control effects on trustor behavior but no evidence for network control effects on trustor behavior. These findings nicely correspond to results from survey research on trust problems in buyer-supplier relations [1, 5, 23]. This survey research focuses on how embeddedness affects trustfulness of buyers in the sense of investing less in costly contractual safeguards that mitigate bad performance, including opportunistic behavior, of suppliers such as delivery of inferior quality, delivery delays, or bad service. Also, this survey research focuses on how embeddedness affects supplier performance itself and thus also supplier trustworthiness. Results indicate that suppliers react to network control opportunities of buyers in the sense that more such control opportunities for buyers are associated with better supplier performance [23]. This finding is nicely in line with our experimental result that trustworthiness increases with network control opportunities of trustors. On the other hand, dyadic control opportunities of buyers do affect their investments in costly contractual safeguards [1], but there is hardly any empirical evidence for effects of network control opportunities on buyer behavior [5]. These findings are in line with our experimental results on effects of control opportunities on trustor behavior.

Buskens (see [5, pp. 152-161]) provides various arguments and also some empirical evidence that the lack of effects of network control opportunities on buyer behavior could be at least partly due to design, data, or measurement problems of the survey, including problems due to possible endogeneity of network embeddedness characteristics and sample selectivity. However, these are no plausible arguments for the lack of network control effects on trustor behavior in our experiment. Thus, one

might wonder whether the findings for effects of control opportunities through embeddedness indicate limits of strategic rationality. First, consider the situation of the trustee (or, respectively, the supplier). He has a good reason to react to the trustor's dyadic control opportunities *as well as* her network control opportunities when he anticipates that his present trustworthiness might affect future trustfulness of the same or other trustors. Similarly, the trustor has a good reason to react to her dyadic control opportunities when she anticipates that the trustee anticipates on how his present trustworthiness will affect this trustor's *own* future trustfulness. However, the trustor needs to reason "more steps ahead" before having a good reason to react to her network control opportunities. Namely, she has to anticipate that the trustee anticipates on how his present trustworthiness will affect future trustfulness of *other* trustors and that *other* trustors will in fact condition their trustfulness on the trustee's present trustworthiness. It may be less likely that actors reason so many steps ahead, certainly in rather unfamiliar settings such as our experiment. Future research could further explore this conjecture in various ways. For example, if the conjecture is correct, we would expect that effects of network control opportunities on trustor behavior are more easily found when trustors play repeated Trust Games with information exchange between trustors many times and specifically when they are also in the role of the trustee in some of those repeated games.

Finally, note that network embeddedness can and, in our experiment, empirically does affect trustfulness of trustors even if the behavior of trustors themselves is not directly affected by their network control opportunities. Since trustees react to network control opportunities of trustors, network embeddedness increases trustworthiness. Through learning effects on trustor behavior, network embeddedness then also increases trustfulness.

Acknowledgment. We thank Davide Barrera, Chris Snijders, and Jeroen Weesie for helpful suggestions and discussion. This article is part of the project "Third-Party Effects in Cooperation Problems" funded by the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Utrecht University High Potential-program of Vincent Buskens and Stephanie Rosenkranz "Dynamics of Cooperation, Networks, and Institutions." Raub acknowledges funding by the Netherlands Organization for Scientific Research (NWO) for the research projects "The Management of Matches" (S 96-168 and PGS 50-370) and "Commitments and Reciprocity" (400-05-089).

References

1. Batenburg, R.S., Raub, W., Snijders, C.: Contacts and Contracts: Temporal Embeddedness and the Contractual Behavior of Firms. Res. Sociol. Organ. 20, 135−188 (2003)
2. Bolton, G.E., Katok, E., Ockenfels, A.: How Effective Are Online Reputation Mechanisms? An Experimental Study." Manage. Sci. 50, 1587−1602 (2004)
3. Bolton, G.E., Ockenfels, A.: ERC: A Theory of Equity, Reciprocity, and Competition. Am. Econ. Rev. 90, 166−93 (2000)
4. Bower, A., Garber, S., Watson, J.C.: Learning about a Population of Agents and the Evolution of Trust and Cooperation. Int. J. Ind. Organ. 15, 165−190 (1997)
5. Buskens, V.: Trust and Social Networks. Kluwer, Boston (2002)

6. Buskens, V.: Trust in Triads: Effects of Exit, Control, and Learning. Game. Econ. Behav. 42, 235−252 (2003)
7. Buskens, V., Raub, W.: Embedded Trust: Control and Learning. Adv. Group Process. 19, 167−202 (2002)
8. Buskens, V., Raub, W.: Rational Choice Research on Social Dilemmas. Forthcoming in Wittek, R., Snijders, T.A.B., Nee, V. (eds.) Handbook of Rational Choice Social Research. Russell Sage, New York (2008)
9. Buskens, V., Raub, W., Van der Veer, J.: Trust in Triads: An Experimental Study. ISCORE-paper. Utrecht Univ. (2008)
10. Camerer, C.F.: Behavioral Game Theory. Experiments in Strategic Interaction. Russell Sage, New York (2003)
11. Camerer, C.F., Weigelt, K.: Experimental Tests of a Sequential Equilibrium Reputation Model. Econometrica 56, 1−36 (1988)
12. Coleman, J.S.: Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge (1990)
13. Dasgupta, P.: Trust as a Commodity. In Gambetta, D. (ed.) Trust: Making and Breaking Cooperative Relations, pp. 49−72. Blackwell, Oxford (1988)
14. Fehr, E., Schmidt, K.M.: A Theory of Fairness, Competition, and Cooperation. Q. J. Econ. 114, 817−868 (1999)
15. Fischbacher, U. z-Tree - Zurich Toolbox for Readymade Economic Experiments – Experimenter's Manual. Exp. Econ. 10, 171−178 (2007)
16. Fudenberg, D., Levine, D.K.: The Theory of Learning in Games. MIT Press, Cambridge (1998)
17. Granovetter, M.: Economic Action and Social Structure: The Problem of Embeddedness. Am. J. Sociol. 91, 481– 510 (1985)
18. Greiner, B.: The Online Recruitment System ORSEE 2.0. A Guide for the Organization of Experiments in Economics. Univ. of Cologne, Working Paper Series in Economics 10 (2004)
19. Harsanyi, J.C.: Games with Incomplete Information Played by 'Bayesian' Players I-III. Manage. Sci. 14, 159−182, 320−334, 486−502 (1967/68)
20. Kreps, D.M.: Corporate Culture and Economic Theory. In Alt, J.E., Shepsle, K.A. (eds.) Perspectives on Positive Political Economy, pp. 90−143. Cambridge University Press, Cambridge (1990)
21. Rabin, M.: Incorporating Fairness into Game Theory and Economics. Am. Econ. Rev. 83, 1281−302 (1993)
22. Raub, W., Weesie, J.: Reputation and Efficiency in Social Interactions: An Example of Network Effects. Am. J. Sociol. 96, 626–654 (1990)
23. Rooks, G., Raub, W., Tazelaar, F.: Ex Post Problems in Buyer-Supplier Transactions: Effects of Transaction Characteristics, Social Embeddedness, and Contractual Governance. J. Manage. Governance 10, 239−276 (2006)
24. Rusbult, C.E., Van Lange, P.A.M.: Interdependence, Interaction, and Relationships." Annu. Rev. Psychol. 54, 351–375 (2003)
25. Snijders, T.A.B., Bosker, R.J.: Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Sage, Thousand Oaks (1999)
26. Van Lange, P.A.M.: The Pursuit of Joint Outcomes and Equality in Outcomes: An Integrative Model of Social Value Orientation. J. Pers. Soc. Psychol. 77, 337–49 (1999)
27. Yamagishi, T., Yamagishi, M.: Trust and Commitment in the United States and Japan. Motiv. Emotion 18, 129−166 (1994)

Third Party Reputation in Repeated Trust Games

Riccardo Boero[*], Giangiacomo Bravo[°],
Marco Castellani[†], Francesco Laganà[†],
Flaminio Squazzoni[†]

[*] Dipartimento di Scienze Economiche e Finanziarie "G. Prato", Università di Torino and GECS - Research Group in Experimental and Computational Sociology, Università di Brescia, boero@econ.unito.it
[°] Dipartimento di Scienze Sociali, Università di Torino and GECS -Research Group in Experimental and Computational Sociology, Università di Brescia, giangiacomo.bravo@unito.it
[†] Dipartimento di Studi Sociali and GECS — Research Group in Experimental and Computational Sociology, Università di Brescia {gbravo, castella, squazzon}@eco.unibs.it
[†] Dipartimento di Studi Sociali, Università di Brescia, lagana@eco.unibs.it

Abstract. This paper elaborates on some first results of a laboratory experiment on the relevance of reputation for the emergence of cooperation in repeated trust games. We have extended the Keser's repeated trust game (Keser 2003) adding new treatments where reputation is taken more explicitly into account. We compare treatments where the investor and the trustee rate each other and treatments where the investor and the trustee are rated by a third party. The results show that reputation positively affects trust and cooperation but also that some differences in the reputation mechanism can generate different cooperative outcomes.

Keywords: Reputation; Third Party; Trust Game; Laboratory Experiments.

1 Introduction

Recent experimental studies have explicitly acknowledged the relevance of reputation in strengthening trust and cooperation in human societies. A good example is Keser (2003), where reputation is introduced into repeated trust games. In a trust game, the investor (*A* player) decides how much of his/her endowment to send to the trustee (*B* player), who receives the amount sent multiplied by a given factor greater than one (usually three). Subsequently, *B* decides the proportion of the received amount to return to *A* (e.g., Berg *et al.* 1995; Camerer 2003). In Keser's design, subjects interacted for a fixed number of rounds and *A* players were allowed to rate the behavior of their opponents. In the following round, the reputation scores were presented to the new *A* players before they took their decision. The Keser's main result was that the possibility of relying on others' experiences significantly increases cooperation in the game, by providing incentives for *A* investments and promoting *B* returns.

Experimental studies and analytical and simulation models on reciprocity confirmed that the possibility of knowing others' past behavior significantly increases cooperation (e.g., Novak and Sigmund 1998a; 1998b; 2005; Fehr and Fischbacher 2003; Seinen and Schram 2006; Bravo and Tamburino 2008). The point is that recent experimental finding suggested that reputation does not need to be necessarily grounded on direct knowledge of past behavior of other individuals, since humans can exploit the many sources of information available in social life, nor it is just closely related to reciprocity mechanisms. For instance, gossip plays an important role in transmitting reputational information in human societies (e.g., Burt and Knez 1995; Dunbar 1996; Kurubawa 2005; Ferrin, Dirks and Shah 2006). Some recent works showed that gossip has a strong influence on the behavior of experimental subjects even when subjects can use other information sources, e.g. direct

observation (Sommerfeld *et al*. 2007). Another experiment showed that individuals react to the possibility of being the subject of gossip by increasing their contributions in a dictator game (Piazza and Bering 2008). The normative nature of reputation mechanisms has also been emphasized in many simulation models (e.g., Raub and Weesie 1990; Conte & Paolucci 2002; Janssen 2006; Hahn *et al*. 2007).

This paper aims to show some first experimental results where the role of third parties in reputation mechanisms is taken explicitly into account. In a previous paper, we have extended the Keser's repeated trust game (Keser 2003) by allowing both *A* and *B* players to rate their opponents and by making reputation scores available for other players in following rounds. Results shown that human subjects are highly sensitive to their own reputation and react promptly to other people's judgments. Although a large part of the reputation effect was due to a strategic investment in reputation made by players (e.g., Bolton, Katok and Ockenfels 2004; Barrera 2008), our results shown that reputation matters even when any practical effect of reputation scores on subjects' payoffs was ruled out. Such results help corroborate the hypothesis that subjects attach to reputation a sound normative flavor (Boero *et al*. 2008).

In this paper we present four new treatments that further extend the reputation-based repeated trust game presented in Boero *et al*. (2008) so as to introduce a third party reputation mechanism. In these new treatments, reputation scores are not directly expressed by the same players (*A* and *B*) who are engaged in interaction as in the previous treatments, but by *C* players who do not take part in interaction and are called to observe and rate *A*'s and *B*'s behavior. We aim to focus on two reputation mechanisms. In the first one, reputation scores are assigned and transmitted by *C* players. In the second one, reputation scores of *B* players are assigned by *A* players and transmitted to *C* players who decide whether transmit the scores to the other players. In this case, *C* players do not have full responsibility on scores' assignment, so that reputation mechanism looks more like gossip. Our experiments help focus on the following research questions: Does a third party reputation mechanism generate the same cooperative outcomes that direct reputation can guarantee in a trust game? What is the consequence of a gossip-like third party reputation mechanism on trust and cooperative outcomes?

The paper is organized as follows. In the second section we describe the experiment. The third one presents the results. The fourth one suggests concluding remarks and future developments.

2. The Experiment

The experiment took place on two days in the computer laboratory of the Faculty of Economics of the University of Brescia, which is equipped with the experimental software z-Tree (Fischbacher 1999). The subjects were 126 students of the University of Brescia recruited through public announcements in the various faculties. All interactions took place through the computer network and the subjects were unable to identify their counterparts. They played 25 rounds of our reputation-based repeated trust game. The subjects were informed in advance of the duration of the game. Each subject played for nearly one hour and the average earning was about 15 Euros, which was paid immediately after the experiment.

The experiment followed four previous treatments where we extended the Keser's repeated trust game to allow *A* and *B* players to rate their opponents (Boero *et al*. 2008). First, we replicated the repeated trust game baseline, which is as follows: (i) subjects are randomly assigned to the role of *A* and *B* players each round; (ii) both player *A* (the investor) and player *B* (the trustee) received an initial endowment of 10 experimental currency units (ECU), having an exchange rate of 1 ECU = 1.5 Euro cents; (iii) player *A* decided his/her investment and the invested amount was tripled and sent to player *B* in addition to his/her own endowment; (iv) *B* chose the amount to return to *A*; (v) the sums earned by both players in the current period were displayed to both subjects. Upon the

baseline, we add three new treatments. In the first treatment (indicated as *Treatment 1*), *A* was allowed to rate *B's* behavior as "negative", "neutral", or "positive", as in Keser's repeated trust game. The subsequent *A* players interacting with *B* were informed of the last reputation score received by the latter *before* making their investment decision. The second treatment (*Treatment 2*) was exactly like the first except for the fact that only *B* was allowed to rate *A*. This information was available in the next period for the subject playing as *B* with *A* players who were already rated. As in previous cases, the possible scores were "negative", "neutral" and "positive" and this information was presented to the *B* players before their return decision. In the third treatment (*Treatment 3*), both *A* an *B* players were allowed to rate each other and knew the scores in the following period before their investment/return decision. The results of this first experiment are used in this paper as baselines against which new treatments are compared.

The new experiment consisted of four new treatments (called Treatment *C1, C2, C3* and *Inf*). Subjects are randomly assigned to the roles of *A, B* and *C* players each round. *Treatment C1* worked as *treatment 1* with the exception that *B's* behavior is rated as "negative", "neutral", or "positive" by *C* players who are called to observe the interaction. Reputation scores of *B* players are transmitted by *C* players to the subsequent *A* players. In treatment *C2, C* players observed and rated *A's* behavior. Reputation scores are transmitted to the subsequent *B* players. In *treatment C3, C* players observed and rated both *A's* and *B's* behavior and reputation scores are transmitted to both. *Treatment Inf* worked as *treatment C1* with the exception that *B's* reputation is not assigned by *C* players but by *A* players. Scores are then transmitted to *C* players who decide whether to transmit the scores to subsequent *B* players. Before to decide, *C* players know the past reputation score of the *A* players the last time they played in the *B's* role. While the difference between *treatment 1, 2* and *3* and *treatment C1, C2* and *C3* is that in the second case reputation scores originated from a player that was not directly involved in the interaction (third party), the difference between *C1* and *Inf* is that, in the second case, *C* players are simply *informers* who do not have any direct responsibility on the quality of reputational information they transmit. As a matter of fact, in real social life, it is likely that reputation takes place also when people relate to reputational information at second hand, such as in case of gossip (e.g., Conte and Paolucci 2002). *Treatment Inf* allows us to observe the impact of a gossip-like reputation mechanism on trust and cooperation.

3. Results

Let us begin with the results of treatments *1, 2, 3* and *C1, C2*, and *C3*. To get a reasonable comparison, treatments should be matched as follows: *1* with *C1* (*B* players are under rating), *2* with *C2* (*A* players are under rating), 3 with C3 (both *A B* players are under rating). While in *treatment 1, 2* and *3* reputation took place within the groups of players, in *C1, C2* and *C3* reputation was due to a third party that was external to the groups of players. Table 1 show mean and standard deviation of *A* investments, *B* returns and B return proportional to the *B*'s endowment.

In *treatment 1* and *C1*, before to decide whether and how much to invest, *A* can use the *B*'s reputation score that another *A* has assigned to *B* at the end of the previous round (*treatment 1*), or the *B*'s reputation score assigned by C, that is, a third party (*treatment C1*). *B* returns are important to understand whether *B* is interested in building a good reputation to *A's* eyes. Results show that the introduction of a third party basically replicates the high level of cooperation of *treatment 1*. *A* investments and *B* returns are even higher in absolute terms, although *B* returns in proportion to the endowment are essentially the same. Some statistical analyses on the ratio between proportional returns and rating in *treatment 1* and *C1* show that there is a co-graduation between the treatments. *Treatment 1* shows a value of γ=0,591 and Kendall's τ=0,481 and *treatment C1* shows corresponding values of γ=0,589 and Kendall's τ=0,464. This means that in the two treatments subjects essentially behaved in the same way. Moreover, the non significance of Mann-Whitney test

shows that the distribution of B returns are not significantly different between the two treatments (p=0.709).

In *treatment 2* and *C2*, before to decide whether and how much to return to *A*, *B* can use the *A's* reputation score. In this case, *B* can make use of his/her direct evidence (how much *A* has decided to invest) and the past *A*'s reputation as well, with this second information source that should not make great sense if *B* players would behave as 'pure' rational agents. The results show that *C* players essentially do not make a real difference as regards to *treatment C2*. *A* players invested slightly more in *treatment 2* than in *treatment C2* (see Table 1), but the difference is not statistically relevant (see Table 2). The same holds true when we look at the absolute (*p=0.086*) and proportional (*p=0.572*) returns in *treatment 3* and *C3* (see Table 2). Although some small difference can be pointed out, such as a higher investment of *A* players in *treatment 3* (*p=0.01*), the *treatment 3* and *C3* show the same results, with the same *B* proportional returns. This is confirmed when we look at the Mann-Whitney test comparison among these six treatments in Table 2. The same trend holds true when we look at the average of *B's* returns and the consequent reputation score assigned by *A* at the end of each round (see table 4), that is, when reputation score's assignment takes place.

Table 1. Mean and standard deviation of *A* investments, *B* returns and *B* returns proportional to the endowment.

Treatments	*Statistics*	*A investment*	*B return*	*B return*
1	Mean	4.29	6.40	0.25
	Std. Dev.	2.70	4.84	0.14
2	Mean	5.61	5.83	0.19
	Std. Dev.	2.77	5.80	0.17
3	Mean	5.27	7.08	0.24
	Std. Dev.	3.02	5.98	0.16
C1	Mean	5.00	6.70	0.24
	Std. Dev.	3.39	6.16	0.18
C2	Mean	5.46	4.86	0.16
	Std. Dev.	3.37	5.88	0.17
C3	Mean	4.64	6.59	0.24
	Std. Dev.	3.08	6.59	0.20

Table 2. Mann-Whitney test comparison of *treatment 1-C1*, *2-C2*, and *3-C3* (one-tailed).

Treatments	*A investment*	*B return*	*B proportional return*
1-C1	W=27742	W=32108.5	W=32562.5
	p=0.010	*p=0.6455*	*p=0.2578*
2-C2	W=32099	W=35357.5	W=35265.5
	p=0.3563	*p=0.0082*	*p=0.0097*
3-C3	W=35017.5	W=33719	W=32421.5
	p=0.01	*p=0.08653*	*P=0.286*

Table 3. Average of *A* investments, *B* returns and *B* proportional in *treatment 1* and *C1*, *2* and *C2*, and *3* and *C3* for past reputation scores of players.

	Treatment 1		*Treatment C1*	

B's score	*A investment*	*B return*	*B proportional return*	*A investment*	*B return*	*B proportional return*
Unknown	4,12	5,66	0,24	4,19	5,81	0,22
Negative	2,57	3,34	0,17	4,41	4,70	0,18
Neutral	4,03	6,21	0,27	4,32	5,52	0,22
Positive	6,33	10,00	0,33	6,16	9,41	0,31
A's score		*Treatment 2*			*Treatment C2*	
Unknown	6,18	5,15	0,52	3,19	4,55	0,45
Negative	3,25	3,91	0,39	3,38	4,05	0,41
Neutral	4,44	4,91	0,49	5,10	5,15	0,52
Positive	7,79	6,99	0,70	6,49	7,03	0,70
B's score		*Treatment 3*			*Treatment C3*	
Unknown	5,46	6,46	0,22	4,29	5,71	0,22
Negative	4,32	4,07	0,15	3,76	3,11	0,13
Neutral	5,67	8,33	0,28	4,49	5,89	0,22
Positive	5,82	9,30	0,31	5,45	9,55	0,33
A's score						
Unknown	8,55	5,37	0,54	5,66	4,14	0,41
Negative	6,84	4,65	0,47	5,24	3,38	0,34
Neutral	6,04	4,23	0,42	5,93	4,82	0,48
Positive	7,38	6,14	0,61	8,51	5,78	0,58

Table 4. Average *B's* return and reputation score assigned by *A*/*C* players after the *B's* return at the end of each round in *treatment 1* and *C1*.

		Treatment 1		*Treatment C1*	
Score assigned to B	*Statistics*	B returns	B proportional returns	B returns	B proportional returns
Negative	Mean	3.47	.1289	3.48	.1226
	Std. Dev.	3.85	.1387	4.45	.1333
Neutral	Mean	6.57	.2743	5.55	.2088
	Std. Dev.	4.19	.0887	4.89	.1523
Positive	Mean	9.21	.3645	9.27	.3588
	Std. Dev.	4.51	.1087	6.12	.1693

Figure 1 and 2 show the dynamics of *A* investments and *B* returns over time. Results indicate that the third party reputation mechanism guaranteed on average a higher level of cooperation at the beginning of the game (see Figure 1 and 2 on the right). Cooperation does not increase or tends even to decline as the rounds follow. Looking at the average *A* investments, *B* returns and *B* proportional in *treatment 1* and *C1* (see Table 3), where the target of reputation scores were *B* players, we can argue that investments/returns and reputation scores have had a strong relation. This is confirmed when looking at values of Kendall's τ=0,481 (*treatment 1*), τ=0.5962 (*treatment 2*), τ=0.5984 (*treatment 3*, proportional investment), τ=0.5670 (*treatment 3*, proportional return), τ=0.4647 (*treatment C1*), τ=0.5317 (*treatment C2*), τ=0.4560 (*treatment C3*, proportional investment), τ=0.4176 (*treatment C3*, proportional return), τ=0.5009 (*treatment Inf*).

Fig. 1. *A* investments over time. From left top, *treatment 1* and *C1*, *2* and *C2*, and *3* and *C3*.

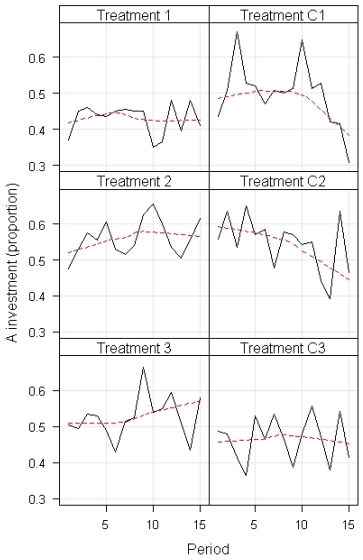
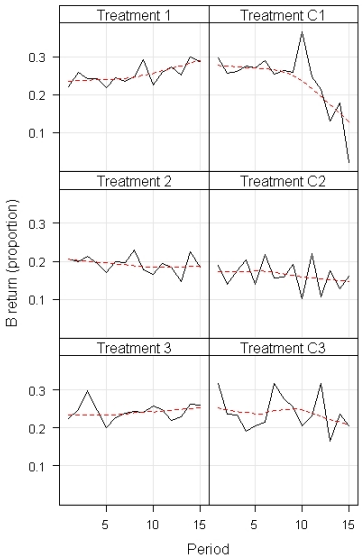


Fig. 2. *B* returns over time. From left top, *treatment 1* and *C1*, *2* and *C2*, and *3* and *C3*.



The result of *treatment Inf* shown that A invested on average 4.16 ECU and B returned on average 4.36 ECU, and 0.17 in proportion to his/her endowment, that is, less that in *treatment 1* and *C1*. The same evidence holds true when we look at Mann-Whitney test that compares *treatment Inf* with *treatment 1* and *C1*, and *treatment 1* with *treatment Inf* (see Table 5). *C* players decided to send to *A* players the *B's* reputation score in 67% of the cases. When *B* players had a negative reputation score, *C* players decided to pass it along in 62% of the cases, when a good one, in 72% of the cases, when a neutral one, in 60% of the case, when unknown, in 81% of the cases. The reputation score that *A* players assigned to *B* players followed these *B's* returns and proportional returns: *A* assigned a negative score when *B* returned on average 2.26 ECU (proportional return is .089), a neutral score when *B*

returned on average 4.44 ECU (proportional return is .198), and a positive when *B* returned on average 9.27 ECU (proportional return is .320).

No significant effect has been played also by past reputation scores of *A* players when they played in a *B's* role. *C* players decided to send the *B's* reputation score when *A* had a negative reputation in 48% of the cases, a neutral in 21% of the cases and a positive in 31% of the cases. They decided not to pass the *B's* reputation score when *A* had a negative reputation in 52% of the cases, a neutral in 24% of the cases and a positive in 34% of the cases. The past reputation score of *A* players did not have great meaning to *C* players.

These last evidences allow us to conclude that the decision of the *informers* as such did not significantly impact the game. But, if this is true, what else can explain the relatively worst performance of this last reputation mechanism as regards to what happened in *treatment 1* (direct reputation) and *C1* (third party reputation)?

Table 5. Mann-Whitney test comparison of *treatment 1-C1, 2-C2,* and *3-C3* (one-tailed).

Treatments	*A investment*	*B return*	*B return (proportion to B's endowment)*
1-*Inf*	W=32769 *p=0.2177*	W=40901 *p=0.000*	W=41295 *p=0.000*
C1- *Inf*	W=18958.5 *p=0.006*	W=16862.5 *p=0.000*	W=17006.5 *p=0.000*

Figure 3 shows the dynamics of *B's* returns over time in *treatment Inf*. The evidence is that, at the beginning of the game, players started by cooperating but soon in a few rounds cooperation started to decline according to a clear pattern. The fact that players started by cooperating means that they understood the game and did not approach it following a negative flavor. Patterns of players' behavior are relatively stable and not affected by random factors.

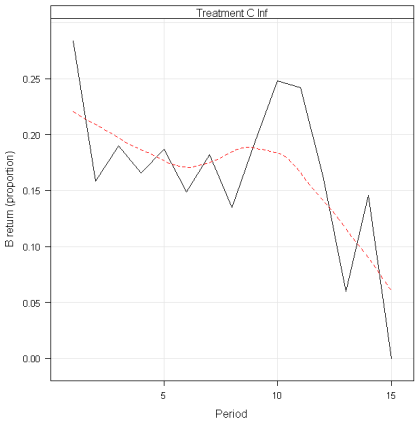


Fig. 3. *B's* returns over time in *treatment Inf*.

To capture this pattern, we have dissected the impact of reputation scores in *treatment 1*, *C1* and *Inf* (see Figure 4, 5 and 6). The comparison shows that *treatment Inf* is characterized by the predominance and the persistence of negative scores. At the end of the round 15, 69% of *B* players have had negative reputation scores, 51% of whom kept a negative score nearly along all the game (see Figure 6). A possible explanation is that A players started to cooperate and to invest, but B players returned less than in *treatment 1* and *C1*. This, in turn, brought about negative reputation scores for *B* players, reduction of investments of *A* players and weak incentives for the emergence of trust and cooperation.

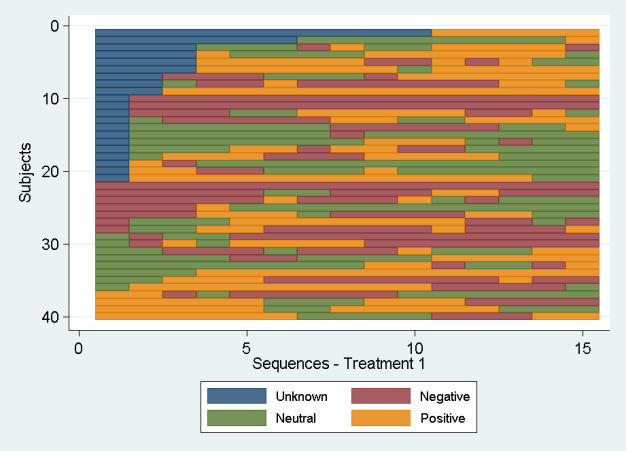


Fig. 4. Sequence of *B's* ratings in *treatment 1*.

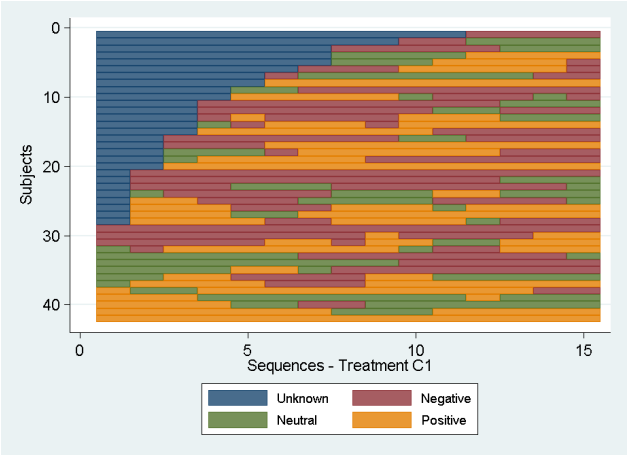


Fig. 5. Sequence of *B's* ratings in *treatment C1*.

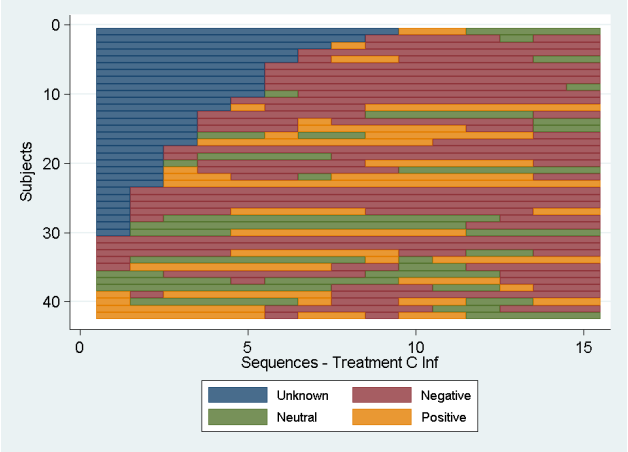


Fig. 6. Sequence of *B's* ratings in *treatment Inf*.

There are two possible explanatory mechanisms of *B's* behavior, which could be at work simultaneously in a complementary way. First, *B* players decided to return little because they could suspect to be cheated by *C* informers, who could decide not to pass along their possible positive reputation to other *A* players. Secondly, *B* players decided to return little

because they thought that *C* informers did not have any concrete incentive to pass to *A* players their possible negative reputation. The point is that *C* informers have been viewed by players as a source of uncertainty in the reputation process. Players thought that *C* informers were not responsible for the quality of the reputational information they were called to spread, because scores were expressed by the players themselves. On the contrary, the introduction of the third party reputation mechanism in *treatment C1* was viewed as a stable, largely predictable reputation mechanism, exactly because in that case *C* players were responsible for the score they were called to pass along. The evidence is that our gossip-like reputation mechanism did not guarantee the same level of trust and cooperative regimes that was guaranteed by the third party reputation mechanism, where *C* players called to assign reputation had full responsibility of the reputation they handed out.

4. Concluding Remarks

The experiments presented in this paper have shown two main findings that can help extend the explanatory power of reputation-based trust for the emergence of cooperation among humans. First, our experiments have shown that human subjects tend to be really sensitive to their reputation even when this would not affect their material payoffs. This means that subjects did not take into account only material self-interest consequences of their behavior, but mostly referred to social norms that are tied to the quality of the intentions and behavior of others (e.g., Rabin 1993). Subject in our experiments have shown the capacity of spontaneously converging on implicit normative scaffolds that have guided their decision in strategic interactions. Reputation has allowed subjects to "encapsulate" trust, to paraphrase Cook, Hardin and Levi (2005). Secondly, the effect of third party reputation mechanism on trust and cooperation can be positive provided that third parties are perceived by other players to be responsible of the reputational information that they transmit. This can explain the bad performance of treatment *Inf* when compared with *treatment* 1 and *C1*, where reputation was based on reputation scores expressed directly by the same players or by *C* players who had the full responsibility of assigning the score.

Future developments include, first, the deepening of the analysis of the experimental data presented in a preliminary stage in this paper, trying to further dissect the impact of reputation mechanisms in all the treatments. In particular, our future aim is to map the sequence of behavior of players at a more fine-grained observation level, so as to capture interaction mechanisms in more detail. Secondly, we would like to test other treatments where the role of the informers can be more clarified than at the present. Our experimental results show that the presence of informers as reputation carrier has caused higher uncertainty for players. The consequence is that, in *treatment Inf*, there is little trace of the normative scaffolds that were spontaneously achieved by players in the other treatments. By providing different information and strongest incentives to *C* players when they play as informers, we could try to further investigate the difference between reputation transmitted in third party relations and gossip transmitted by informers as carriers of trust and cooperation.

Acknowledgements

Financial support was provided by a FIRB 2003 grant (Strategic Program on Human, Economic and Social Sciences) from the Italian Ministry for the University and the Scientific Research (SOCRATE Research Project, coordinated by Rosaria Conte, Protocol: RBNE03Y338_002). We would like to thank three anonymous conference referees for very helpful remarks and comments. Usual disclaimers apply.

References

Barrera, D.: The social mechanisms of trust. Sociologica, 2, doi: 10.2383/27728 (2008)

Berg, J., Dickhaut, J., & McCabe, K. A.: Trust, reciprocity and social history. Games and Economic Behavior, 10, 122-142 (1995)

Boero, R, Bravo, G., Castellani, M., Squazzoni F.: Reputational clues in repeated trust games. University of Brescia Department of Social Sciences Working Paper SOC 01-08: http://www.unibs.it/on-line/dss/Home/Inevidenza/PaperdelDipartimento/documento8952.html, Submitted to Journal of Socio-Economics (2008)

Bolton, G. E., Katok, E., and Ockenfels, A.: How effective are electronic reputation mechanisms? An Experimental investigation. Management Science, 50, 1587-1602 (2004)

Bravo, G. and Tamburino, L.: The evolution of trust in non-simultaneous exchange situations. Rationality and Society, 20, 1, 85-113 (2008)

Burt, R. and Knez, M.: Kinds of third-party effects on trust. Rationality and Society, 7, 3, 255-292 (1995)

Camerer, C. F.: Behavioral game theory. Experiments in strategic interaction. New York/Princeton: Russell Sage Foundation/Princeton University Press (2003)

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

Cook, C., Hardin, R., and Levi, M.: Cooperation without trust?. New York: Russell Sage Foundation (2005)

Dunbar, R. I. M.: Grooming, Gossip and the evolution of language. Cambridge, MA: Harvard University Press (1996)

Fehr, E. and Fischbacher, U.: The nature of human altruism. Nature, 525, 785-791 (2003)

Ferrin, D. L., Dirks, K. T. and Shah, P. P.: Direct and indirect effects of third-party relationships on interpersonal trust. Journal of Applied Psychology, 91, 4, 970-883 (2006)

Fischbacher, U.: z-Tree. Zurich toolbox for readymade economic experiments. University of Zurich, Working paper no. 21 (1999)

Hahn, C., Fley, B., Florian, M., Spresny, D., and Fischer, K.: Social reputation: A mechanism for flexible self-regulation in multiagent systems. JASSS. 10(1): <http://jasss.soc.surrey.ac.uk/10/1/2.html> (2007)

Janssen, M.: Evolution of cooperation when feedback to reputation scores is voluntary. JASSS. 9(1): <http://jasss.soc.surrey.ac.uk/9/1/17.html> (2006)

Keser, C.: Experimental games for the design of reputation management systems. IBM Systems Journal, 42, 498-506 (2003)

Kuwabara, K.: Affective attachment in electronic markets. A sociological study of eBay. In Nee V. and Swedberg R. (Eds.), The economic sociology of capitalism, Princeton, Princeton University Press, 268-288 (2005)

Nowak, M. A. and Sigmund, K.: The dynamics of indirect reciprocity. Journal of Theoretical Biology, 194, 561-574 (1998a)

Nowak, M. A. and Sigmund, K.: Evolution of indirect reciprocity by image scoring. Nature, 393, 573-577 (1998b)

Nowak, M. A. and Sigmund, K.: Evolution of indirect reciprocity. Nature, 437, 1291-1298 (2005)

Piazza, J. and Bering, J. M.: Concerns about reputation via gossip promote generous allocations in an economic game. Evolution and Human Behavior, 29, 172–178 (2008)

Rabin, M.: Incorporating fairness into game theory. The American Economic Review, 83, 1281-1302 (1993)

Raub, W. and Weesie, J.: Reputation and efficiency in social interaction: An example of network effect. American Journal of Sociology, 3, 626-654 (1990)

Seinen, I. and Schram, A.: Status and group norms: Indirect reciprocity in a helping experiment. European Economic Review, 50, 581-602 (2006)

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., and Milinski, M.: Gossip as an alternative for direct observation in games of indirect reciprocity. PNAS 104(44), 17435-17440 (2007)

The Impact of Seller Reputation on Simultaneous Auctions of Identical Goods: Theory and Experimental Evidence

Extended Abstract[1]

Ravi Bapna, University of Minnesota
Chrysanthos Dellarocas, Robert H. Smith School of Business, University of Maryland
Sarah Rice, The University of Connecticut

Contact email: cdell@rhsmith.umd.edu

Introduction and Motivation

While much of auction theory and online auction research has focused on the analysis of single auctions, online auction marketplaces typically consist of multiple competing auctions offering identical or very similar goods sold by sellers of different reputations. As an illustration of this fact, if, on September 1, 2008, an aspiring iPod owner searched eBay for an "iPod nano 2GB", she would have been confronted with 1,710 practically identical listings sold by sellers of widely varying eBay feedback scores and reputation profiles. Our hypothetical bidder only needed one iPod, raising the question as to whose auction(s) should she place her bid(s) and in what amount(s)? The objective of this work is to provide rigorous answers to this common, but surprisingly under-researched scenario. There is an emerging theory around competing auctions that draws on some of the early work on sequential auctions (Ashenfelter, 1989) and some recent work on auctions with many traders (Peters and Severinov, 2006) but no results on how unit demand bidders should bid in simultaneous auction settings by sellers of varying reputation. There is, similarly, a significant body of empirical literature on the impact of seller reputation on bidder behavior and seller revenue in settings where there is only one seller (see Dellarocas 2006 and Resnick et al. (2005) for a comprehensive review) but no studies of how reputation affects bidder behavior in auction settings with multiple sellers. Analytical and experimental work documenting the effects of reputation on economic decisions includes that of Kreps (1982), Wilson (1985), Berg, Dickaut and McCabe (1995), Lucking-Reiley (2000) and Bolton et al. (2004).

The vast majority of the online auction reputation literature is based on observational empirical settings that use some variant of hedonic regression techniques to tease out the impact of reputation on variables of interest. However, the impact of differing seller reputation in a simultaneous auction setting, the likely default setting facing most consumers on sites such as eBay, has not yet been studied. Consider, for instance, the main simultaneous auction prediction of Peters and Severinov (2006). They posit that for competing auctions, the final price of one auction is affected by the existence of other auctions, and that prices are expected to be uniform across competing auctions. However, can we expect uniform prices under reputational difference of sellers? We bridge this gap by deriving equilibrium bidding strategies in simultaneous auctions of identical goods by sellers of different reputations. We further test the predictions of this theory in a controlled laboratory setting with economic incentives (Smith 1976). Much like Resnick et al (2005), our motivation for using the controlled laboratory setting is to avoid possible omitted variable bias from the presences of other covariates that impact reputation and auction outcomes, albeit in more complex setting of simultaneous auctions. We also benefit from measuring subject's

[1] This work consists of a theoretical piece and an experimental study. The theoretical piece is complete and reported fully at http://ssrn.com/abstract=1115805. The experimental piece is work-in-progress that will be completed before the end of 2008.

risk profiles through a post experiment survey and analyzing their micro-level bidding behavior captured by our custom designed online auction platform.

Our research questions are: 1) How does seller reputation affect bidding strategy in simultaneous auctions for identical goods? 2) How does observed bidding behavior compare to theoretical predictions in this setting? 3) How do individual risk preferences affect bidding behavior in these types of markets? We report on data from the first round of experiments as well as preliminary findings from our risk assessment survey. Additional data collection is planned for October 2008.

Our experimental analysis finds interesting conformance as well as some interesting departures from theoretical predictions. We find that 78% of bidders resort to bidding on a single auction and (with only one exception) the remaining bidders place two bids. In contrast, theory predicts that all bidders should place multiple bids. Bidders who bid on two auctions, tend to primarily adopt a "hi-low" strategy that overemphasizes the top and (surprisingly) *the bottom* rated sellers. Bidders targeting lower reputation (higher risk) sellers generally overbid, failing to appropriately adjust for the associated risk. Taken together the last two effects result in economic rents for the highest *and lowest* reputation sellers and lower bidder surplus than theory would predict. However this tendency ameliorates over time and we observe market efficiency increases as rounds progress and bidders learn to adjust for the appropriate risks.

Summary of Theoretical Model[2]

Dellarocas (2008) models a setting where a number of sellers of different reputations simultaneously offer sealed-bid, second-price, single-unit auctions of identical goods to unit-demand buyers. A seller's reputation is defined as the buyers' common subjective belief that the seller will fulfill the transaction. The buyers' decision problem is to select on which auctions to submit bids as well as how much to bid. Our study compares the predicted Bayes-Nash bidding equilibria, expected seller revenue, buyer surplus, and allocative efficiency of Dellarocas (2008) to observations derived from student subjects in a controlled induced value laboratory experiment designed as per Smith(1976), Roth (1988), and Kagel and Levin (1993).

Below we provide a qualitative summary of the theoretical results of Dellarocas (2008) that frame our experiment. The analysis distinguishes between two settings: One where bidders are restricted to submit at most one bid and one where bidders have no such restriction. When bidders are restricted to submit at most one bid, there exists a unique Bayes-Nash equilibrium that has the following form: Sellers are ranked according to their reputation. Buyers self-select into a finite number of zones, according to their types (private valuations for the good). Buyers whose types fall in the highest zone always bid on the highest seller; buyers whose types fall in the 2nd highest zone randomize between the top two sellers, assigning higher probability to selecting the 2nd seller; more generally, buyers whose types fall in the k-th zone randomize between the top k sellers, assigning increasingly higher probability to selecting less reputable sellers. This bidding behavior is a form of probabilistic positive assortative matching: bidders assess where they stand on the valuation scale and assign higher probability to bidding on the auction that ``matches" their respective zone, while also occasionally bidding on ``higher" auctions. This strategy is depicted in Figure 1.

[2] Full details of the theoretical model are available at http://ssrn.com/abstract=1115805

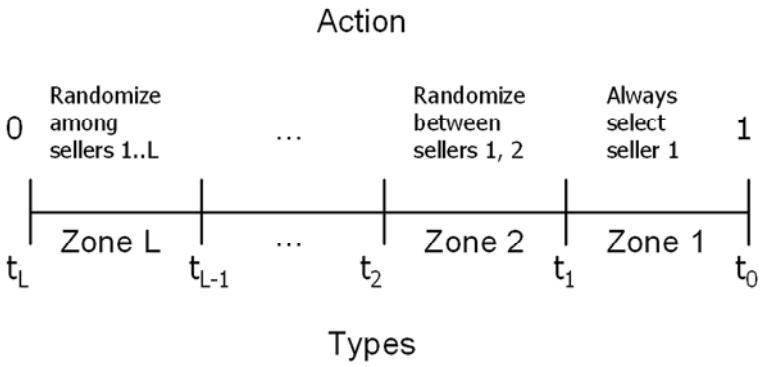


Figure 1- Bidders' Equilibrium Strategy Resembles Probabilistic Assortative Matching

The probabilistic nature of the matching between bidders and sellers creates allocative inefficiencies that are most severe when the number of bidders is roughly equal to the number of sellers. To appreciate this, compare Figure 1 with a centralized assignment problem that a social welfare maximizing social planner could enforce. This important drawback of independent simultaneous auctions could, in theory at least, be remedied by combining single-unit auctions that end at roughly the same time into a single multi-item, provably efficient auction, such as a VCG auction.

In terms of seller revenue, the key property is that more reputable sellers attract higher valuation bidders with higher probability. Reputation, thus, has a double impact on a seller's expected revenue: First, it affects the amount that buyers are willing to bid (it is equal to the valuation of the good times the seller's reputation). Second, it affects the quality of buyers who are more likely to choose a seller's auction. In settings with competition a seller's expected revenue is, then, a convex function of his reputation.

The analysis is subsequently extended to the more general setting where buyers are allowed to bid on an arbitrary number of simultaneous auctions. Surprisingly, theory then predicts that even though they have unit demand, all bidders would place non-zero bids *on all* auctions. The optimal bid amount in each auction is equal to the bidder's expected valuation of the respective auction (taking into account the seller's reputation) multiplied by the probability that the bidder will not receive the item from any of the other auctions, given her other bids. Equilibrium bid vectors of all except the lowest bidder types have one of two forms: I) a high bid (i.e. a bid that is close to the bidder's expected valuation) on one auction and low bids on all remaining auctions or II) intermediate bids on all auctions. Bidders randomize between using Type I and Type II bid vectors. When using Type I vectors they further randomize with respect to which auction they place their high bid on. Higher bidder types use Type I bid vectors more often and place their high bid on higher auctions with higher probability.

Laboratory Experiment

Using these theoretical results to motivate our laboratory experiment we employ economic incentives drawing from the work of Vernon Smith (1976) on induced value theory to control for heterogeneous values and study the decision making behavior of bidders in a controlled setting. We design an auction market where there are six bidders, each with a private value drawn from a

uniform distribution with support (6, 10). Bidders have the option to bid from four seller types, each with a different reputation rating (100%, 90%, 80%, 70%). The reputation score indicates the probability that the buyer will receive the purchased good as advertised. This means that if a bidder wins an auction from a seller with a 90% reputation score the probability that the winner will receive their good is .9; however, regardless of whether or not the good "ships" the winner must always pay for the unit won. If the bidder wins two units he/she must pay for both even though their demand is for one unit only. In each round bidders may bid on as many seller auctions as they choose. Once bidders have submitted bids a subsequent screen indicates whether they have won the good and shows the profit in that round. The experiment lasts for twenty rounds. Subjects are paid in cash at the end of the experiment and their total profits are calculated as follows:

*Profit = [20 rounds * (Value of Good – Price Paid if Winner)] + Participation Fee*

In the above expression *Value of Good* is equal to the bidder's private value if the good is received or zero if it is not.

Our auction setting has an element of risk, in that buying from a seller with anything other than a 100% reputation score could lead to lost revenue if the good is paid for but not shipped. As reputation scores decrease the risk that a bidder would pay for a good but not receive it increases, thus we consider the risk type of subjects as having a possible effect on bidding behavior. To test this we deploy a post experimental risk assessment tool developed and validated by Weber, Blais and Betz (2006). This survey allows us to make inferences about bidding behavior based on risk type and allows greater insight into observed outcomes.

Results

Bidding Strategy: We first analyze whether bidders conform to the predicted bidding strategy; results are shown in Table 1. We observe that 78% of the time bidders bid on a single seller, 21% of the time they bid on two sellers and only once was there an instance of a bidder placing bids on 3 sellers. Because bidders did not predominantly follow the multi-bid strategy we compare our results to the theoretical single bid setting. Our data supports the prediction that higher bidder types are more likely to go for higher reputation sellers, whereas the low types target the low reputation sellers. As predicted and evident in Table 1 lower type (valuation) bidders tend to place multiple bids more often than higher type. We observe that the *bid-to-valuation ratio*[3] is higher in the case of one bid than it is in the case of two bids, suggesting that, consistent with theory, when people place multiple bids they scale at least one of them down relative to their expected valuation. Very often the second bid is a "lowball" bid. Moreover, our data indicates relatively high bid-to-valuation ratios for the highest and lowest reputed seller. This is an interesting observation as it appears that bidders are unable to appropriately adjust for the high risks associated with the 70% reputation sellers.

[3] The *bid-to-valuation ratio* is defined as (bid amount)/((private value)*(seller reputation)) and reflects the degree to which bidders bid their expected valuation for a seller's good, taking into consideration the non-fulfillment risk that is embodied in that seller's reputation. Our theory predicts that if bidders place only one bid the b-t-v ratio should always be equal to 1.

# of simultaneous bids	# of cases	Average bidder value	Average bid-to-valuation ratio
0	9	6.35	0
1	166	8.05	.86
2	45	7.74	.73
3	1	9.07	.63

Table 1 – Bidders Conform to Predicted Equilibrium Bidding Strategy

Bidder Surplus: We find that a higher than theoretically expected proportion of bids were placed with the highest *and lowest* reputation sellers. In addition, bidders tended to over bid on the low reputation sellers, placing bids in excess of their expected value[4]. It is possible that these results reflect risk attitudes which can be explained by prospect theory (Kahneman and Tversky 1979), where decisions appear risk averse in the domain of gains and risk seeking in the domain of losses. Another possible explanation is that individuals have a utility for maximizing profits as well as winning the auction, in which case their utility function might resemble: *V = max(profit) + max(prob(win auction)).* In this instance bidders would bid an additional marginal amount to increase the probability of winning the auction. We find that all bidder types at some point bid on the low reputation seller, therefore it is likely that bids placed by high value bidders' crowd out low value bidders while increasing seller revenue. Overall, the crowding of bidders on the two extreme sellers and the overbidding on the lowest seller result in bidder surplus that is roughly 60%-70% of what theory would predict (Figure 2).

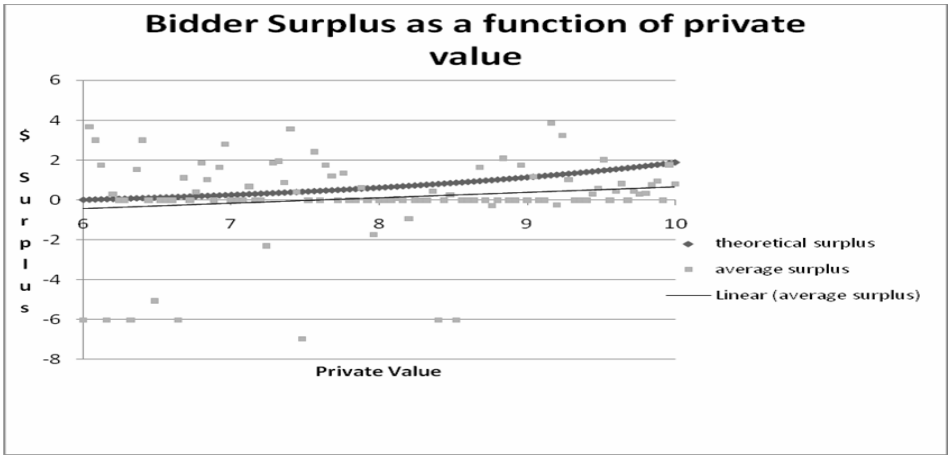


Figure 2

Seller Revenue: We find that average seller revenue is very close to what theory predicts for the two middle sellers but about 30% higher than what theory predicts for the top and

[4] Expected value = (bidder's private value) * (seller's reputation)

bottom seller. Again, this observation can be explained by more bidders bidding on the highest and lowest sellers and bidding higher than expected value in order to increase the probability of winning the auction.

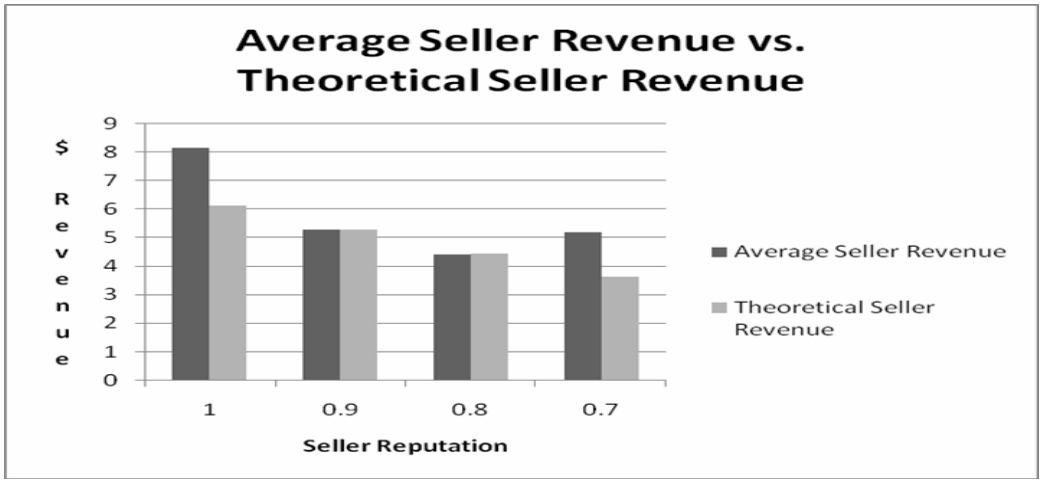


Figure 3

Allocative efficiency: Our theoretical efficiency is .80 and we find that our experimental results are quite close to this level. We also find that average efficiency increases per round suggesting that bidders learn how to bid over time. Without a coordination mechanism such as a Vickery auction or Groves Clark auction it is unlikely that full efficiency is attained as bidders do not know for certain where they fall in the distribution of values, resulting in uncertainty about which seller to bid on. We posit that a solution to the efficiency loss would be to impose a coordination mechanism such as Vickery Groves Clark to generate the efficient allocation.

Impact of risk attitudes: Multivariate analysis of our risk assessment survey shows that bidders who had a higher level of risk tolerance (more willing to take financial risks) were more likely to bid with lower reputation sellers and to bid in multiple auctions. We find that less risk tolerant bidders were likely to bid in only one auction and with the higher reputation sellers. Additional analysis is required to determine if these risk preferences aid in explaining deviations from our predicted outcomes.

Conclusion

Despite the increasing practical importance of reputation in multi-unit auction markets, there is relatively little theory on how unit-demand bidders should behave in this setting. In addition to informing buyer behavior, such research will help sellers better understand how their reputation and the condition of their items affect their expected revenue in the presence of competition. Finally, this work will assist market operators design more effective auction and reputation mechanisms for such environments.

Our results are preliminary and we are in the process of collecting additional data. Specifically we will run a series of treatments with two very low value sellers (30%, 50%) and two much higher value sellers (90%, 100%) to test our theoretical finding that reputations below a certain threshold have little or no value. If accepted we will look forward to presenting the full set of data at ICORE 2009.

References

Ashenfelter, O. (1989). How Auctions Work for Wine and Art. *Journal of Economic Perspectives*. Vol 3. No. 3.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior* 10, pp. 122-142.

Bolton, G., Katok, E. Ockenfels, A. (2004) How Effective are Electronic Reputation Mechanisms? *Management Science*. Vol 50.

Dellarocas, C. (2006). Reputation Mechanisms. *Handbook on Economics and Information Systems* (T. Hendershott, ed.), Elsevier Publishing.

Dellarocas, C. (2008) Simultaneous Auctions of Imperfect Substitute Goods by Sellers of Different Reputations. *Robert H. Smith School Research Paper No. RHS 06-057. Available on SSRN.*

Kagel, J. and Levin D. (1993). Independent Private Value Auctions: Bidder Behavior in First- Second- andThird-_ Price Auctions With Varying Number of Bidders. *The Economic Journal*. Vol. 103. No. 419.

Krishna, V., and R. W. Rosenthal (1996). Simultaneous Auctions with Synergies. *Games and Economic Behavior* 17, 1-31.

Kreps, D. Wilson, R. (1982). Reputation and Imperfect Information. *Journal of Economic Theory.*

Kahneman, D. Tversky, A. (1979). Prospect Theory. *Econometrica*.

Lucking-Reiley, D. (2000). Auctions on the Internet: What's Being Auctioned, and How? *Journal of Industrial Economics* vol. 48, no. 3, pp. 227-252.

Peters, M. and Severinov, S. (1997) Competition Among Sellers Who Offer Auctions Instead of Prices. *Journal of Economic Theory*, 75, 141-79.

Resnick, P., Zeckhauser, R., Swanson, J., K. Lockwood. (2005), The Value of Reputation on eBay: A Controlled Experiment, forthcoming *Experimental Economics*.

Roth, A. E. (1988). Laboratory Experimentation in Economics: A Methodological Overview. *The Economic Journal*, Vol. 98, No. 393

Smith, Vernon L. *(*1976*),* Experimental Economics: Induced Value Theory, *American Economic Review*, 66: 2, pp. 274-279.

Weber, E. Blais, R. Betz, N. (2002). A Domain-Specific Risk-Attitude Scale: Measuring Risk Perceptions and Risk Behaviors. *Journal of Behavioral Decision Making.* Vol 15.

4 Simulation of Reputation

Simulating the Human Factor in Reputation Management Systems for P2P Networks

Guido Boella, Marco Remondino, Gianluca Tornese

Dipartimento di Informatica. Università di Torino - IT.
E-mail: {guido,remond}@di.unito.it; gianluca.tornese@libero.it

Abstract. A compelling problem in peer-to-peer (P2P) networks for file sharing, is the spreading of inauthentic files. To counter this, reputation management systems (RMS) [1] have been introduced. These systems dynamically assign to the users a reputation value, which is considered in the decision to download files from them or not. RMS are proven, via simulation, to make P2P networks safe from attacks by malicious peer spreading inauthentic files. But in large networks of millions of users non-malicious users get a benefit from sharing inauthentic files due to the credit system. In this paper we show using agent based simulation that reputation systems are effective only if there is a widespread cooperation by users in verifying authenticity of files starting during the download phase, while the size punishment derived by the reputation systems is less relevant. This was not evident in previous works since they make several ideal assumptions about the behavior of peers who have to verify files to discover inauthentic ones. Agent based simulation allows to study the human factor behind the behavior of peers, in particular the advantage of spreading inauthentic files, of not checking as soon as possible their authenticity during the download, thus unwillingly cooperating to the spreading of files.

1 Introduction

One of the most compelling problems in peer-to-peer (P2P) networks for file sharing, at least from the point of view of users, is the spreading of inauthentic files; since multimedia files are usually quite large, downloading the wrong ones causes waste of time, of storage space and of bandwidth. The percentage of inauthentic files circulating over a P2P network is high, especially for those resources which are recent and most requested. Some of the reasons of this problem are the possibility to attack a P2P networks introducing malicious peers and how users exploits the protocol with which the current P2P systems reward the users for uploading files, no matter if they are authentic or not. Since uploading bandwidth is a limited resource and the download priority queues are based on a uploading-credit system to reward the most collaborative peers on the network, some malicious users, if they see that a resource is rare or most wanted, could decide to create an inauthentic file for it, just to have something to share, thus obtaining credits. In this way inauthentic files spread very quickly on the network and those malicious users are never penalized for their behavior.

To balance the incentive for sharing inauthentic files, reputation management systems have been introduced. These systems gather information from users about the

authenticity of files downloaded by peers and based on this assign dynamically a reputation to them, which is used as a value in the decision to download files from them or not. Reputation management systems are proven, via simulation, to make P2P networks safe from attacks by malicious peer, even when forming coalitions.

If attack to P2P network via inauthentic files is a relevant problem, it is not the only one. In particular, in large networks of millions of users attacks are more difficult but users still have a benefit from sharing inauthentic files. Reputation management systems have been proved useful against attacks but it is not clear if they can be effective also against widespread selfish misbehavior. The reason is that they make many ideal assumptions about the behavior of peers who have to verify files to discover inauthentic files: this operation is assumed to be automatic and to have no cost. Moreover, since in current systems files are shared before downloading is completed, peers downloading inauthentic files involuntarily spread inauthentic files if they are not cooperative enough to verify their download as soon as possible.

Thus, in this paper we address the following research questions:

- Which are the factors controlled by users which determine the correct functioning of reputation management systems in P2P network?
- How to evaluate the role of these factors by means of agent based simulation? Which factors have most impact?

In the present work the creation and spreading of inauthentic files is not considered as an attack, but a way in which some agents try to raise their credits, while not possessing the real resource that's being searched by others. A basic and idealized reputation management system (RMS) is introduced, acting as a positive or negative reward for the users. We are not interested in simulating factors like distribution of workload among the peers, how reputation calculated by peers, bandwidth concerns and multiple sources downloading, since our focus is different. Instead, the human factor behind the RMS is considered, like the decision of creating fake files in which circumstance and costs and benefits of verifying files.

To verify the limits and effectiveness of a reputation mechanism under different user behaviors we use agent based simulation. A multi agent simulation of a P2P network is used as methodology, employing reactive agents to model the users.

This paper is structured as follows. First in Section 2 we discuss reputation management systems, then in Section 3 we present the P2P scenario we use. Section 4 analysis the parameters and results of simulations. Conclusions discuss the results and future work.

2 Reputation

It is well known that selfish peers wish to use file sharing services while contributing minimal or no resources themselves: to minimize their cost in bandwidth and CPU utilization freeholders refuse to share files in the network. To incentively sharing of files credit systems are set up to give more bandwidth or priorities in waiting queues to users who share more files.

The credit mechanism, alas, can be exploited by users who try to get more credits by distributing highly requested files even if they do not have them, it is sufficient to replace them with inauthentic copies. This behavior can take the form of an explicit creation of inauthentic files or the voluntarily or not voluntarily sharing of downloaded (or still in download) inauthentic files. When a user is downloading a file after a given amount of time he can check its authenticity. If he does not verify, or he does and does not remove that file, from the perspective of the system he becomes a malicious user too.

Reputation management systems, like EigenTrust [2], try to alleviate the problem of inauthentic files by allowing users to gave feedback about the cooperativity of other users and the system uses this feedback to compute a reputation of the different peers in the systems.

In general, a reputation system assists agents in choosing a reliable peer (if possible) to transact with when one or more have offered the agent a service or resource. To provide this function, a reputation system collects information on the transactional behavior of each peer (information gathering), scores and ranks the peers based on expected reliability (scoring and ranking), and allows the system to take actions against malicious peers while rewarding contributors (response).

The driving force behind reputation system design is providing a service that severely mitigates misbehavior while imposing a minimal cost on the well-behaved users. To that end, it is important to understand the requirements imposed on system design by each of the following: among them the behavior and expectations of typical good users, and the goals and attacks of adversaries.

A system must be accessible to its intended users, provides the level of functionality they require and does not hinder or burden them to the point of driving them away. Therefore, it is important to anticipate any allowable user behavior and meet their needs, regardless of added system complexity.

Usability of a systems consist both in providing an acceptable level of service, and in not demanding them costly behaviors, or at least providing incentives for such behaviors.

However, note that malicious behavior by peer is due not only by the will to attack the system or to get higher credits, but also by the cost of verifying the authenticity of files: the user needs a correct version of the P2P software, has to find a suitable viewer, to select a not yet verified file, to wait until it opens, to check the content (sometimes an embarrassing task) and to signal the inauthenticity of the file. Since controlling the authenticity of files is costly, and this operation is a prerequisite for giving a feedback on the cooperativity of peers, the reputation management system is not granted to function as expected.

Most approaches, most notably EigenTrust [2], assume that verification is made automatically upon the start of download of the file. By looking as we do at the human factor in dealing with RMS, we can question their real applicability, a question which remains unanswered in the simulation based tests made by the authors.

To provide an answer to this question it is necessary to build a simulation tool which aims at a more accurate modelling of the users' behavior rather than at modelling the reputation system in detail.

3 The scenario

3.1 The P2P model

The P2P network is modeled as an undirected and non-reflexive graph. An undirected graph is one in which the lines do not have direction. In reflexive graphs, it is possible for a node to have a tie with itself, which is called a loop or, redundantly, a self-loop. In non-reflexive graphs, such links are excluded.

Each node is an agent, representing a P2P user. The agents are reactive: their behavior is thus determined a priori, and the strategies are the result of the stimuli coming from the environment and of the condition-action rules. Their behavior is illustrated in next section.

Formally the multi agent system is defined as $MAS =< Ag, Rel >$, with Ag set of nodes and Rel set of edges. Each edge among two nodes is a link among the agents and is indicated by the tuple $< a_i, a_j >$ with a_i and $a_j \in Ag$.

Each agent features the following parameters:

- Unique ID (identifier),
- Reputation value (or credits) $N(a_i)$,
- Set of agent's neighbors $RP(a_i)$,
- Set of owned resources $RO(a_i)$,
- Set of goals (resource identifiers) $RD(a_i)$,
- Set of resources being downloaded $P(a_i)$,
- Set of pairs $< supplier, resource >$.

A resource is a tuple $< Name, Authenticity >$, where $Name$ is the resource identifier and $Authenticity$ is a Boolean attribute indicating whether the resource is authentic or not. The agent owning the resource, however, does not have access to this attribute unless he verifies the file.

The resources represent the object being shared on the P2P network (e.g.: multimedia files and so on). A certain number of resources are introduced in the system at the beginning of the simulation; they represent both the owned objects and the agents' goals. For coherence, an owned resource can't be also a goal, for the same agent. The distribution of the resource is stochastic. During the simulation, other resources are stochastically introduced in the pool of the agents. In this way, each agent in the system has the same probabilities to own a resource, independently from her inner nature (malicious or loyal).

In the same way also the corresponding new goals are distributed to the agents; the only difference is that the distribution probability is constrained by its being possessed by an agent. Formally R be the set of all the resources in the system. We have that $RD(a_i) \subseteq R$, $RO(a_i) \subseteq R$ and $RD(a_i) \cap RO(a_i) = \emptyset$.

Each agent in the system features a set of neighbors $N(a_i)$, containing all the agents to which she is directly linked in the graph: $N(a_i) = \{a_j \in Ag \mid< ai, aj >\in Rel\}$. This information characterizes the information of each agent about the environment. The implemented protocol is a totally distributed one, so looking for the resource is heavily based on the set of neighbors.

In the real word the shared resources often have big dimensions; after finding the resource, a lot of time is usually required for the complete download. In order to simulate this the set of the "resources being downloaded" (Ris) introduced. These are described as $Ris =< resourceID, completion, checkstatus >$, where ID is the resource identifier, completion is the percentage already downloaded and "check status" indicates whether the resource has been checked for authenticity or not. In particular, it can be not yet verified, verified and authentic and verified and inauthentic:
$check\_status \in \{NOT\_CHECKED, AUTHENTIC, INAUTHENTIC\}$

Another relevant information is ID of the provider of a certain resource, identified by $P(a_i)$. Each agent keeps track of those which are uploading to him, and this information is preserved also after the download is finished. In the real world, the P2P client should keep this information and link it to the received files, in the form of an association provider ID-resource ID. In the model described in this work this information a resource features just a provider for all the downloading process. The real P2P systems allow the same resource to be download in parallel from many providers, to improve the performance and to split the bandwidth load. This simplification should not affect the aggregate result of the simulation, since the negative payoff would reach more agents instead of just one (so the case with multiple provider is a sub-case of that with a single provider).

3.2 The reputation model

In this work we assume a simple idealized model of reputation, since the objective is not to prove the effectiveness of a particular reputation algorithm but to study the effect of users' behavior on a reputation system. We use a centralized system which assumes the correctness of information provided by users, e.g., it is not possible to give an evaluation of a user with whom there was no interaction. The reason is that we focus on the behavior of common agents and not on hackers who attack the system by manipulating the code of the peer application.

3.3 The user model

We model peers as reactive agents replying to requests for download, performing requests or verifying files. While upload is performed each time another agent makes a request, requesting downloading and verification are performed (in alternative) when it is the turn of the agent in the simulation.

All agents belong to two disjoint classes: malicious agents and loyal agents. The classes have different behaviors concerning uploading, while they have the same behavior concerning downloading and verification: as we said, malicious agents are just common agents who exploit for selfishness the weaknesses of the system.

When it is the turn of another peer, and he requests a file to the agent, he has to decide whether to comply with the request and to decide how to comply with it.

- The decision to upload a file is based on the reputation of the requester: if it is below a threshold, the "replying threshold", the requestee denies the upload (even if the requestee is a malicious agent producing inauthentic files).

– The ”replyTo” method refers to the reply each agent gives when asked for a resource. In the reactive model there are two categories for the agents: ”loyal” and ”malicious”. When the agent is faced with a request he cannot comply but the requester's reputation is above the ”replying threshold”, if he belongs to the malicious class, he had to decide whether to create and upload an inauthentic file by copying and renaming one of his other resources. The decision is based depending on a parameter. If the resource is owned, sends it to the requesting agent, after verifying if her reputation is higher than the ”replying threshold”.

Each agent performs at each round of simulation two steps:

1. Performing the downloadings in progress. The first action performed by each agent, for each download in progress, is that of carrying it on; for each resource being downloaded, the agents check if the download is finished (i.e.: it reached n/n parts). If not, in order to avoid inconsistency, the system checks if the resource is still present in the provider's ”sharing pool”. In case it's no longer there, the download is stopped and can (must??) be removed from the list of the ”owned resources”.
   Each file is formed by n units; when 2/n of the file has been downloaded, then the file gets automatically owned and shared also by the agent that is downloading it.
2. Making new requests to other peers or verifying the authenticity of a file downloaded or in downloading, but not both:
   (a) When searching for a resource all the agents within a depth of 3 from the requesting agent are considered (according to the TTL system of Gnutella network, since the network is fixed and thus it is irrelevant to propagate the query each time). The list of these agents is ordered by reputation. A method called ”replyTo” is invoked on every agent with a reputation higher than the ”requests threshold”, till when the resource is found or the list reaches the ending point. If the resource is found, it's inserted in the ”downloading list”, the goal is cancelled, the supplier is recorded and linked with that specific download in progress and her reputation is increased according to the value defined in the simulation parameters. If no resource is found, the goal is given up. However, new goals are introduced in the simulation, according to the parameters set by the user.
   (b) Verification means that a file is previewed and if the content does not correspond to its description or filename, this fact is notified to the reputation system. Verification phase to be carried on, at least one download must be in progress and it must be beyond the 2/n threshold described above. If these hypotheses are satisfied, in the present model an agent has, depending on a parameter, a given probability to check or to look for a new resource and starting a new download if that is found.
   In case the agent decides to verify, a random resource is selected among those being downloaded which has not yet been checked and that's over 2/n of completion. If the resource is authentic, the agent's turn is over. Otherwise, if the resource turns out to be inauthentic, a ”punishment” method is invoked, the resource is deleted from the ”downloading” list and from the ”owned resources” list and the same resource is inserted among the ”goals” once again.

The reputation mechanism is based on the ”punishment” method which lowers the supplier's reputation, deletes her from the ”providers” list in order to avoid cyclic (and multiple) punishment chains, and recursively invokes the ”punishment” method on the punished provider. In this way a punishment chain is created, reaching the creator of the inauthentic file, and all the aware or unaware agents that contributed in spreading the resource.

4 Experimental results

4.1 The variables considered

At present, the simulation goes on until at least one goal (for at least one agent) exists and/or a download is still in progress.

A ”come-back” mode is implemented and selectable before the simulation starts, in order to simulate the real behavior of some P2P users who, realizing that they cannot download anymore (since they have low credits or, in this case, bad reputation), disconnect their client, and then connect again, so to start from the initial pool of credits/reputation. When this mode is active, at the beginning of each turn all the agents that are under a given threshold reset it to the initial value, metaphorically representing the disconnection and reconnection. If the starting value for the reputation is properly selected, this practice doesn't necessarily privilege (advantage) these users: in fact the reputation should rise more with a normal file sharing behavior, thus allowing the agents to achieve their goals in shorter times.

In the following table a summary of the most important parameters for the experiments are given:

Parameters	Value
Total number of agents	50
Total number of graph edges	80
Initial reputation (credits) for each agent	50
Percentage of loyal agents	5
Total number of resources at the beginning	50
Threshold to share owned resources	25
Threshold for requesting resources	10
Number of turns for introduction of new resources	1
Number of new resources to introduce	3
Total number of steps	2000

Table 1.

4.2 Simulation methodology

In this section the results of the simulation are depicted and analyzed. In all the experiments, the other relevant parameters are fixed, while the following ones change:

Parameters	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
positive payoff	1	1	1	1	1	1
negative payoff	3	3	4	8	0	2
verification percentage	30%	40%	30%	30%	30%	40%

Table 2.

A very important index, defining the wellbeing of the whole P2P system, is the ratio among the number of inauthentic resources and the total number of resources on the network. As stated before, the total number is increasing more and more over time, since we introduce a number of new resources after a given period. In particular, in the experiment, it has been chosen to introduce three new resources at every time step: this is quite realistic, considering a total number of agents of 50. After 2000 time-steps, which is the limit we choose to use for our observations, we can expect to have a total of 5997 unique resources, summed to the initial 50 (a grand total of 6047). Another measure collected is the average reputation of loyal and malicious agents at the end of the simulation; in an ideal world, we expect that malicious ones to be penalized for their behavior, and good ones to be rewarded. This multi-run technique, described in [3], is a way to structurally validate the model and overcome the sampling effect; the proportion among malicious and loyal agents, the distribution of both kinds of agents in the network, the links among them, the distribution of the resources and goals are all stochastic variables, so running many times a simulation with the same parameters gives an high level of confidence for the produced results. The results were obtained by a batch execution mode for the simulation. This executes 50 times the simulation with the same parameters, sampling the inauthentic/total ratio every 50 steps. In 2000 turns, we have a total of 40 samples. After all the executions have finished, the average for each time step is calculated, and represented in a chart. In the same way, the grand average of the average reputations for loyal and malicious agents is calculated, and represented in a bar chart. In figure 1, the chart with the trend of inauthentic/total resources is represented for the results coming from experiments 1, 2, 3, 5 and 6. The results of experiment 4 will be discussed later.

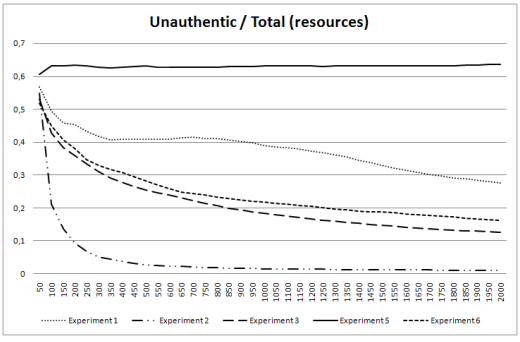


Fig. 1.

4.3 Evaluation of results

Experiment 5 depicts the worst scenario, in which no negative payoff is given for the discovery of inauthentic files. The ratio initially grows and, at a certain point, it gets constant over time. This is explainable by the fact that the new resources are stochastically distributed among all the agents with the same probability: in this way also malicious agents have new resources to share, and they will send out inauthentic files only for those resources they do not own. In the idealized world modeled in this simulation, since agents are 50% malicious and 50% loyal, and since the ones with higher reputation are preferred when asking for a file, its straightforward that malicious agents reputation fly away, and that an high percentage of files in the system are inauthentic (about 63%). Experiment 1 shows how a simple RMS, with quite a light punishing factor (3) is already sufficient to lower the percentage of inauthentic files in the network over time. We can see a positive trend, reaching about 28% after 2000 time steps, which is an over 100% improvement compared to the situation in which there was no punishment for inauthentic files. In this experiment the verification percentage is at 30%, meaning that only 30% of downloaded files are checked for authenticity (during or after their download). This is quite low, since it means that 70% of the files remain unchecked forever (meaning that they have been downloaded, but never used). In order to show how much the human factor can influence the way in which a RMS works, when applied to a P2P system, in experiment 2 the verification percentage has been increased up to 40%, while leaving the negative payoff still at 3. The result is surprisingly good: the ratio among inauthentic resources and the total number of files on the network is dramatically lowered after few turns (less than 10% after 200), reaching less than 1% after 2000 steps. Since 40% of files checked is quite a realistic percentage for a P2P user, this empirically proves that even the very simple RMS proposed in this work dramatically helps in reducing the number of inauthentic files on a P2P network. Later on, also the average reputation of the agents will be discussed, in order to see whether loyal agents also get penalized or not, when acting as unwilling accomplice. In order to assign a quantitative weight to the human factor, in the next experiment, number 3, the negative payoff is taken from 3 to 4, while bringing back the verification percentage to 30%. Even with a higher punishing factor, the studied ratio is worse than it was in experiment 2, meaning that its preferable to have a higher verification rate, compared to a higher negative payoff. In this experiment, anyway, its shown that, with a 30% of verification, the ratio of inauthentic files over total resources goes from 28% (experiment 1) down to about 12.5%, just by moving the negative payoff from 3 to 4. Experiment 6 shows the opposite trend: the negative payoff is lighter (2), but the verification rate is again at 40%, as in experiment 2. The trend is very similar just a bit worse - to that of experiment 3. In particular, the ratio of inauthentic files, after 2000 turns, is about 16%. At this point, it gets quite interesting to find the break even point among the punishing factor and the verification rate, in order to reach a very positive aggregate result for the ratio of inauthentic files, as the one obtained in experiment 2. After some empirical simulations, we have that, compared with 40% of verification and 3 negative payoff, if now verification is just at 30%, the negative payoff must be set to a whopping value of 8, in order to get a comparable trend in the ratio of inauthentic files circulating over the network. This is done in experiment 4 and represented in the chart in figure 2.

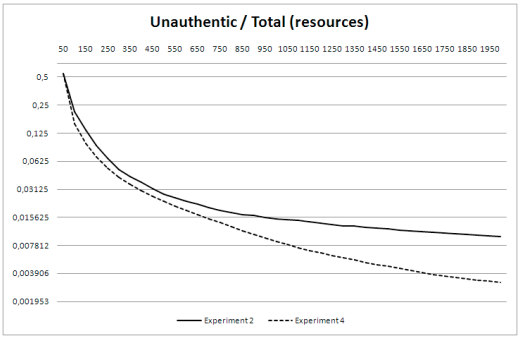


Fig. 2.

The chart uses a logarithmic scale on y-axis, since the trends are almost overlapped: after 2000 turns, theres 1% of inauthentic files with a negative payoff of 3 and a verification percentage of 40%, and about 0.7% with 8 and 30% respectively. This clearly indicates that human factor (the files verification) is crucial for a RMS to work correctly and give the desired aggregate results (few inauthentic files over a P2P network). In particular, a slightly higher verification rate (from 30 to 40%) weights about the same of a heavy upgrade of the punishing factor (from 3 to 8). Besides considering the ratio of inauthentic files moving on a P2P network, its also crucial to verify that the proposed RMS algorithm could punish the agents that maliciously share inauthentic files, without involving too much unwilling accomplices, which are loyal users that unconsciously spread the files created by the former ones. In the agent based simulation, this can be considered by looking at the average reputation of the agents, at the end of the 2000 time steps. This is observed in the bar chart in figure 3, which compares this value for the various experiments discussed before.

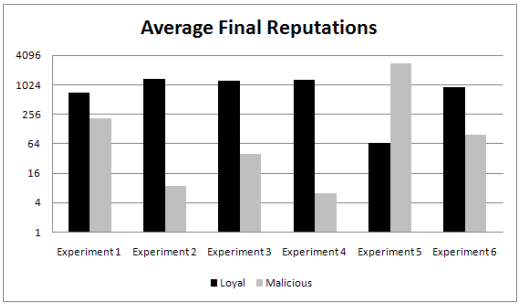


Fig. 3.

The scale for y-axis is logarithmic, since in some experiments the difference among higher and lower values is too high to be visible with a linear one. In the worst case

scenario (with no negative payoff for inauthentic files), depicted in experiment 5, the malicious agents, that are not punished for producing inauthentic files, always upload the file they are asked for (be it authentic or not). In this way, they soon gain credits, topping the loyal ones. Since in the model the users with a higher reputation are preferred when asking files, this phenomenon soon triggers an explosive effects: loyal agents are marginalized and almost ostracized, and never get asked for files. This results in a very low average reputation for loyal agents (around 70 after 2000 turns) and a very high average value for malicious agents (more than 2800) at the same time. In experiment 1 we can see how the basic RMS presented here, can help with this; even with a low negative payoff (3) the average reputations after 2000 turns, the results are clear: about 700 for loyal agents and slightly more than 200 for malicious ones. In the system there are two reputation thresholds: the first and higher one, under which its impossible to ask for resources to other agents, the second, lower than the other, which makes it impossible even to share the owned files. This guarantees that an agents that falls under the first one (because she shared too many inauthentic files), can still regain credits by sharing authentic ones and come back over the first threshold. On the contrary, if she continues sharing inauthentic files, she will fall also under the second threshold, being de facto excluded from the network, still being a working link from and to other agents. In experiment 1, many malicious agents fall under the lower threshold, while no loyal agent does. Thus, the algorithm preserves loyal agents, while punishing malicious ones. As discussed before, in experiment 2 we keep the negative payoff at 3, but higher the verification percentage (human factor) from 30% to 40%. As for the ratio of inauthentic files in the system, this proves to be a tremendous improvement for the effectiveness of the RMS algorithm. In fact, the average reputation for loyal agents, after 2000 steps, reaches almost 1400, while all the malicious agents go under the lower threshold (they cant either download or share resources), with an average reputation of less than 9 points. As for the percentage of inauthentic files in the system, the human factor proves to be a very important aid to the presented RMS; fortunately, 40% of verification is a very likely figure to be carried on in the real word. Experiment 3 explores the scenario in which the users are very uninterested, in the sense that they just check 30% of the files they download, but the negative payoff is raised from 3 to 4. The final figure about average reputations is again very good, even more than it was in experiment 2. Loyal agents, after 2000 steps, averagely reach a reputation of over 1200, while malicious ones stay down at about 40. This again proves the proposed RMS system to be quite effective, though, with a low verification rate, not all the malicious agents get under the lower threshold, even if the negative payoff is 4. In experiment 6 the verification percentage is again at the more realistic 40%, while negative payoff is reduced to 2. Even with this low negative payoff, the results are very good: many (most) malicious agents fall under the lowest threshold, so they cant share files anymore and they get an average reputation of about 100. Loyal agents, on the other side, behave very well and reach an average reputation of more than 900, after 2000 turns. Experiment 4 is the one in which we wanted to harshly penalize inauthentic file sharing (negative payoff is set at 8), while leaving an high laxity in the verification percentage (30%). Unlikely what it could have been expected, this setup does not punish too much loyal agents that, unwillingly, spread unchecked inauthentic files. In fact, after 2000 turns, all the malicious

agents fall under the lowest threshold, and feature an average reputation of less than 7 points, while loyal agents fly at an average of almost 1300 points. The fact that no loyal agent falls under the point of non return (the lowest threshold) is probably due to the fact that they do not systematically share inauthentic files, while malicious agents do. Loyal ones just share the resources that they never check, and which they own, thinking they are the real thing. Malicious agents, on the other side, always send out inauthentic files when they are asked for a resource they do not own for real, thus being hardly punished by the RMS, when the negative payoff is more than 3.

Comeback mode The fact that in the RMS proposed in this work two thresholds exist to prevent malicious agents that persists in producing and sending out inauthentic files from downloading or even sharing resources, makes it straightforward to imagine that an agent could simply disconnect and reconnect to the network, in order to start from an initial reputation value and overcome these limitations. In order to simulate this, a comeback mode has been implemented in the simulation and studied in the following experiments. The agents, when reaching the lowest threshold, are simply reset to the initial reputation value (thats what would happen if they disconnected and reconnected). In Table 3 the parameters for the next experiments are described. All the other parameters stay still as showed in Table 1.

Parameters	Exp 7	Exp 8	Exp 9
positive payoff	1	1	1
negative payoff	3	3	4
verification percentage	40%	30%	30%

Table 3.

In Figure 4 the inauthentic files ratio, resulting from experiments with this mode enabled, is depicted.

It is very interesting to notice that, even with comeback mode activated, the results are very similar to those in which this mode is turned off. They are actually a bit worse when the negative payoff is low (3) and so is the verification percentage (30%): the ratio of inauthentic files in the network is quite high, at about 41% after 2000 turns (experiment 8) versus the 27% observed in experiment 1, which had the same parameters, but no comeback mode active. When the verification percentage is increased to 40%, though, things get quite better, as seen in experiment 7. Now the ratio of inauthentic files has the same levels as in experiment 2 (less than 1% after 2000 steps). Also with a lower verification percentage (again at 30%), but leaving the negative payoff at 4 (experiment 9), the figure is almost identical to the one with the same parameters, but without a comeback mode (experiment 3). In fact, after 2000 turns, the inauthentic files ratio is at about 12%. In figure 5 the bar chart showing the average reputations for the agents in the comeback scenario is examined. The scale on y-axis is logarithmic.

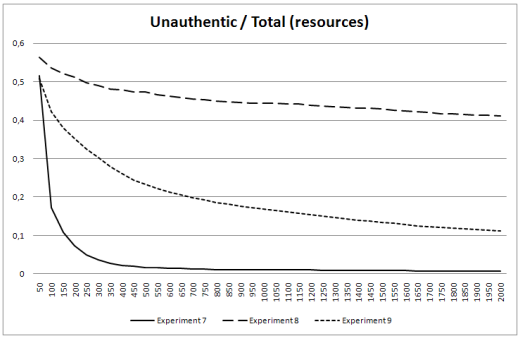


Fig. 4.

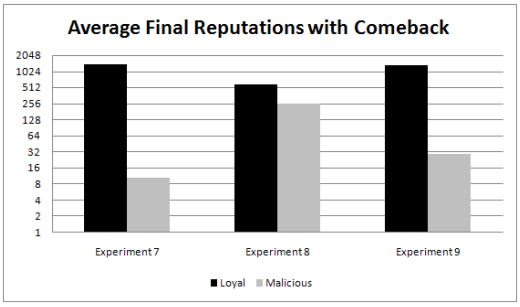


Fig. 5.

Experiment 8, in which both the negative payoff (3) and the verification percentage (30%) are low, the average reputations after 2000 steps are worse than those seen in experiment 1, with the same parameters, but no comeback mode. Loyal agents have now a final average reputation of less than 600, against that of almost 700 seen in experiment 1. Malicious agents have now an average reputation of about 250, while when no comeback was possible this value was barely higher than 200. Experiments 7 and 9, on the contrary, again return values which are almost identical to those seen in the same scenarios with no comeback mode. So, the experiments show that malicious agents, even resetting their own reputation after going below the lowest threshold, cant overcome the presented basic RMS, if they always produce inauthentic files. This happens because, even if they reset their reputation to the initial value, its still quite low compared to the one reached by loyal agents; if they shared authentic files, this value would easily go up in few turns, but since they again start spreading inauthentic files, they almost immediately fall under the thresholds again and again. In future works, a more variable (stochastic and adaptive) behavior will be implemented for the agents, to see how the results change and if the RMS is still reliable in more realistic situations.

Scaling issue : changing the numbers of the agents involved Agent based simulations can suffer from scaling issues, meaning that the results obtained with a certain number of agents can vary when this number changes significantly. In order to examine this, two more experiments were carried on. In those, the number of agents is increased to 150 (thats three times compared to that of the previous experiments). Coherently, the number of edges is also tripled from 80 to 240 and so is the initial pool of resources (from 50 to 150) and the number of resources introduced at each turn (from 3 to 9). The goal is to compare the results from these simulations to those obtained with the former ones, and check if the results are still valid. Experiment 1a is compared with Experiment 1, and experiment 2a with experiment 2, them being the same versions with higher numbers. In figure 6, the comparison among Experiment 1 and 1a is carried on.

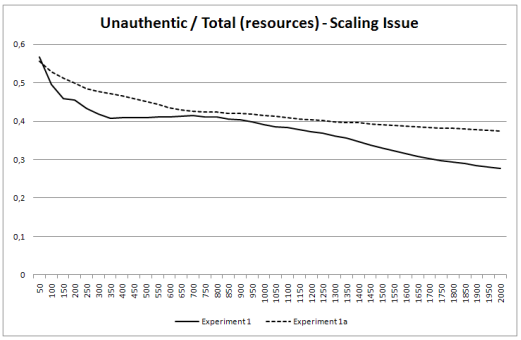


Fig. 6.

The trend is very similar, even if not exactly identical. With a low negative payoff (3) and a low verification rate (30%), a higher number of agents, even if all the other data are increased accordingly, result in a slightly higher ratio of inauthentic files. The same comparison is carried on with the average final reputations, in experiment 1 and 1a. This is shown in figure 7.

Here the results are very similar, even if, again, we see that the system with more agents has a slightly worse aggregate behavior than the smaller one. In figure 8, experiment 2 and 2a are compared. Now the negative payoff is still at 3, while the verification percentage is raised to 40%. In this case, the results are almost identical, so in figure 8 a logarithmic scale was used for y-axis, to be able to see a difference among the lines (which, otherwise, would be overlapping).

Now, increasing the number of agents, the final ratio of inauthentic files on the network is even lower, although almost imperceptibly, when compared to the case where less agents were involved. Lets have a look at the final average reputations of the agents (figure 9). Once again, the logarithmic scale is used for y-axis.

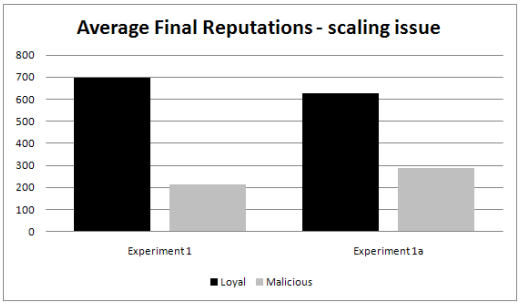


Fig. 7.

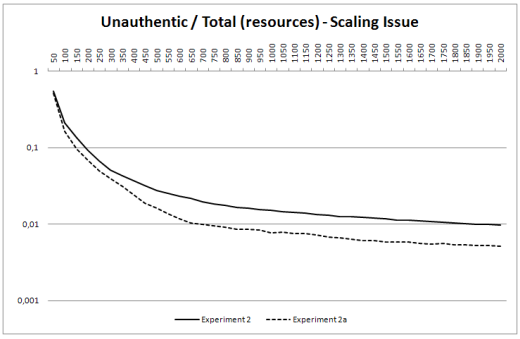


Fig. 8.

Generally we can conclude that the difference is very low, so the scaling issue is not influencing the results shown, at least when moving from 50 to 150 agents. In future works this study will be extended to even more agents.

5 Conclusions

First of all in this paper we question which are the factors controlled by users which determine the correct functioning of reputation management systems in P2P network. We individuate two critical points: the decision of sharing inauthentic files and the decision not to verify the downloaded files.

While the benefit of non verifying is determined by the time saved, since verifying is incompatible with making new searches and starting new download (the simulation scenario model this as a percentage on the possible requests), the benefit of spreading inauthentic files must be confirmed by the simulation. If we run a simulation without the mechanism for punishing malicious agents, inauthentic files will increase sharply, since the peers with highest reputation are malicious agents - and at the end of simulation the reputation of malicious agents is much higher loyal agents', and this difference is

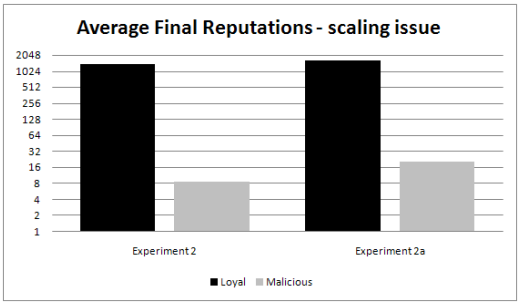


Fig. 9.

reached very soon. Thus, producing inauthentic files is a paying strategy, if there is no enforcement mechanism.

Note however that the behavior of malicious agents strikes back against them, since we assume that they are not attackers and thus they have the goal to download authentic resources too.

In the current system, the behavior of peers, and thus also of malicious agents is fixed, so they cannot adapt their behavior to the flooding of bad files each one contributes to spread.

In this paper we assume that if a file is verified then the reputation of the uploader is decreased immediately, due to the lower cost of this action. A more fine grained model should consider also this human factor. Analogously we do not consider the possibility to punish peer without first receiving and checking a file - a behavior with should be prevented by the software itself - as well as we do not consider the possibility of punishing even if the file is authentic. As stated in the Introduction, our goal is to model the behavior of normal user, not of hackers attacking the system.

The second question of the work is: how to evaluate the role of these factors by means of agent based simulation? Which factors have most impact?

The simulation framework for reputation gives interesting results: the key factor to lower the number of inauthentic files in a file sharing P2P system is the proportion of verifications made by peer. Even a reasonable figure like 30% sharply limits the behavior of malicious agents when we do not consider the possibility of whitewashing after comeback. The value of the punishment in terms of decrease of reputation has instead a more limited impact, in particular in comeback mode

Surprisingly, even when white washing is allowed - thus a malicious agent can disconnect and change identity to joint the system again to repeat his greedy behavior - the number of inauthentic files in the system can be limited if peers verify files 40% of the time they spend making new requests. The same result cannot be achieved by increasing the figure of the punishment and decreasing the proportion of verifications. Under 40%, the system is not able to defend itself from comingback malicious agents.

The moral of our study is that a mechanism for stimulating users to check the authenticity of files should be promoted, otherwise the entire file sharing system is flooded by inauthentic files.

In contrast, most approaches to reputation systems consider verification automatic, thus ignoring the human factor: since we show that verification has a sharp effect which varies according with the proportion it is made by users, it cannot be ignored in simulating the effect of a reputation system.

Thus, we identify the conditions, when even a simple RMS can dramatically reduce the number of inauthentic files over a P2P system and harshly penalize malicious users, without directly banishing them from the network, like proposed in other models based on ostracism, which unrealistically presuppose the possibility of disconnecting a peer at the network level

The model we propose is very simple. The reader must be aware of several limitations, which are the object of ongoing work.

Resources are not divided in categories. Inauthentic files in reality are mostly found in new resources, as common user experience witnesses when downloading new blockbuster movies: authentic files are almost impossible to find. Thus, we are aiming at using real data about download to differentiate the kinds of resources, distinguishing in particular newly requested resources (like when a new movie is distributed in theaters and users try to download it on a P2P network).

We now only distinguish malicious agents from loyal agents, but all agents of each category have the same behavior, for example they verify with the same proportion. It could be useful to simulate what happens when using different parameters in each reactive agent of the two classes.

We do not consider the problem of bandwidth. Thus downloads proceeds all at the same rate, even if the decision to upload to a peer is based on his reputation. Moreover, a peer decides to upload on the basis of which agent has the highest reputation. It is well known that this algorithm risks to create unbalance among peers, but we abstract here from this problem, since we are not proposing a new P2P mechanism but checking the efficacy of a reputation system on the specific problem of inauthentic files. Note, however, that this strategy has a negative effect when malicious peers get high reputation, but if the reputation system is well tuned, malicious agents never get high reputation.

Finally, we allow agents to disconnect and reconnect, but this whitewashing to recover reputation happens without changing position of the agent in the graph, and the reconnected agents behave like before the disconnection.

The real improvement in our ongoing work, however, rests in passing from reactive agents always repeating their behavior to more sophisticated agents able to learn from what is happening in the network. While in the current model agents stochastically decide whether to upload an inauthentic file or not, or to verify or not, it is more realistic that agents adapt to the circumstances, looking how many objectives they can achieve using their current strategy, and looking for new alternatives.

Modeling adaptive agents is important, since it allows to check further vulnerabilities, like what happens when agents produce inauthentic files at a variable rate which does not decrease too much their reputation.

References

1. A. Josang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decis. Support Syst.*, 43(2):618–644, March 2007.

2. S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 640–651, New York, NY, USA, 2003. ACM Press.
3. M. Remondino and G. Correndo. Mabs validation through repeated execution and data mining analisys. *International Journal of Simulation: Systems, Science & Technology*, 7(6), 2006.

From Hazardous Behaviours to a Risk Metric for Reputation Systems in Peer to Peer Networks

Erika Rosas[1] and Xavier Bonnaire[2]

[1] Laboratoire d'Informatique de Paris 6
Université Pierre et Marie Curie - CNRS UMR 7606
Paris - France
[2] Universidad Técnica Federico Santa María
Departamento de Informática
Valparaíso - CHILE

Abstract. Peer to Peer systems have shown to be very powerful to build very large scale distributed information systems. They are self organized, and provide very high availability of the data. However, the management of malicious peers is a very open problem for the Peer to Peer research community, and building trust is a very difficult task.
In this context, Reputation Systems have shown to be a very good solution to build trust in Peer to Peer systems. Nevertheless, using only the reputation value of a peer to decide to make a transaction is not sufficient to guarantee that it will succeed, and the use of the credibility of recommendation emitters does not always significantly mitigate the computed reputation.
We show in this paper the importance of the notion of risk associated to the reputation value, and why a better decision can be taken using both, the reputation and a risk value, for a given peer. We present some metrics based on the list of recommendations for a peer that allow to detect some suspicious behaviours that can alert the application of the presence of a malicious peer. The proposed metric is flexible such that an application can adapt the metric to its needs, given more or less weight to some specific types of behaviours.
We present some simulations to show the influence of malicious behaviours of a peer over its reputation value with the evaluation of the associated risk, and how our metric can detect this kind of behaviours. We conclude about the need to use a risk factor associated to the reputation value, and present some future works about the risk metrics.

1 Introduction

Building trust in Peer to Peer networks is a very difficult task, mainly because of the number of peers, the high dynamism of the network, and the presence of malicious peers. These characteristics make using a certification authority based on a set of servers not a very well suited answer to this problem, as it requires a central administration, which it is not a scalable solution. Other traditional

authentication techniques can not be used because of the ability of the peers to change its identity, and the need of anonymity of the peers [1].

In this context, reputation systems have shown to be a very good solution to build trust in Peer to Peer systems [12], [6], [14], [9], [10]. The key idea of a reputation system is to provide a reputation value for each peer, which can be seen as the probability for the peer to be trusted. To compute the reputation value the system defines a *metric* based on a set of recommendations emitted by other peers after completing a transaction. When a transaction succeeds, a good recommendation must be emitted, and a bad one otherwise. An application can then decide whether or not to do a transaction with a peer according to its reputation value.

Usually, the metric of reputation systems also considers the credibility of the peer which emits the recommendation, as a function of its reputation value [13] [9] [6] or as the similarity of its past evaluations [12] [2]. Nevertheless, the reputation value is not sufficient, and malicious peers can take advantage of a good reputation value to deceive other peers.

As the reputation value is based on the behaviour of the peers, it can not reflect some of the strategies used by the peers to fool the reputation system. This is why the notion of risk has been introduced as a complement to the reputation value. The risk value is used to try to detect suspicious behaviours of the peers that have a good reputation and seem to be trusted. The purpose is to give the reputation system more information, about the analysis of the behaviour of the peers, to make the system more robust to attacks of malicious peers.

To our knowledge, the notion of risk as presented in this paper has never been proposed before. Only the work in Pet [8] reputation system introduces the notion of risk in their trust model. In Pet, this is a value derived from direct interactions with other peers. This is a very different approach since it only take into account a short-term behaviour [8], and it is focused to detect sudden changes of behaviour of the peers that the reputation value can not detect. A drawback of this work is that in peer to peer networks, with millions of peers it is not very probable that a peer had already a previous direct interaction with other specific one.

A few proposals have attempted to address the issue of malicious attacks to the reputation system. Overall, reputation systems are focused on mitigating malicious recommendations, which are detected with the use of a credibility value. Xiong and Liu in [12] consider the problem of free riders adding to the reputation metric a community context factor, which can be a function of the feedback provided by the peer to the reputation system. This is a way to encourage the participation of peers.

TrustGuard [11] is a framework that is focused, as our work, on understanding the vulnerabilities of the reputation systems and on how to minimize the effects of malicious peers. The difference is that TrustGuard changes the reputation metric to achieve this. We believe that the reputation metric gives valuable information itself and can be quite flexible for an application, but we also believe that an application needs additional information to know if the reputation value of a peer

can be trusted itself. TrustGuard [11] detects three vulnerabilities, malicious peers that adapt its behaviour to maximize its malicious goals, rumours and false recommendations. We did not consider in our work the last two problems because they can be mitigated directly in the reputation metric. A solution based on a proof of transaction (evidence) has been proposed in [11]. We will see later on in this paper that our approach of the risk uses an analysis of the behaviour of the peer, based on the list of recommendations that the reputation system already has to calculate the reputation value.

The RQC reputation system [5] proposes a quality function to evaluate the trustworthiness of the reputation value. Similarly to some metrics in our work, they consider the number of recommendations and the variance of the data to compute the quality of the reputation value. RQC searches the consistence in the reputation value more than detect suspicious attacks of malicious peers that take advantage of their reputation value to attack the system.

In this paper, we propose a risk metric capable to detect several well-known malicious behaviours of peers, such that the Oscillating Personality, the Random Behaviour, and the Repeated One Shot Attack.

The rest of the paper is organized as follows. In Sect. 2, we briefly present a general model of a reputation system where a risk metric can be applied. Section 3 details a set of risk metrics to detect several well-known malicious behaviours of peers. Then, experiments and results are shown in Sect. 5. Finally, conclusion and future work are presented in Sect. 6.

2 Reputation System Model

The risk metrics presented in the next section are based in the idea that to compute the reputation value of a peer X ($Re(X)$) the reputation system collects a number of recommendation emitted by some peers which already had transactions with X in the past.

We note $F_i(X)$ the recommendation emitted about a peer X of index i from a total of m recommendations. The value m in some systems can be considered like a sufficient number of recommendations or in others as the maximal number of recommendations to compute a reputation.

We suppose in the following that the reputation value is the probability for a peer to be trusted, and that the reputation system uses recommendations in the range [0..1], with at least three discreet values.

There are several reputation systems that follow this model [9], [12], [2], [13]. All of them could include a risk metric as a complement to the reputation value in order to help an application to decide whether or not to make a transaction.

3 Malicious Behaviours and Associated Risk Metrics

There are several strategies that a malicious peer can use to fool the reputation system. None of them can be detected using only the reputation value of the

peer. An application can then ignore a wrong behaviour of this peer. In this section, we present a set of well-known malicious behaviours for a peer, and we propose an associated risk metric capable of detecting this malicious behaviour.

3.1 White Washers

A peer is called a White Washer when it intentionally leaves the network and enter again with a new identity, in order to clear its history of recommendations. This allows the peer to fool an application, appearing with a fresh good reputation. This is mainly due to the assignment of a good reputation to new peers entering the network (positive discrimination) to give them a chance to make a transaction. Therefore, it becomes difficult to discriminate new peers from malicious ones for the reputation system. The worst case appears in the Sybil Attack [3] where a peer can have multiple identities.

In decentralized reputation systems there are no solutions to identify these peers, but there are some ways to mitigate their impact. The use of expensive identifiers can help to prevent a peer from trying to get several different identifiers, due to the computational or financial cost to obtain a new identifier.

Giving a reputation to the resources used in the network (i.e. files, etc...) like in [2] [7], or giving a low reputation value to new peers can help to mitigate the effects of White Washers. However, this does not encourage new honest peers to participate to the system. The work of Friedman in [4] has shown that the distrust in new peers is a social cost inherent to the easy change of identity.

The problem with the reputation value is that a peer X with a number of good recommendations $r \ll m$, will have a similar reputation value that a peer with m good recommendations. For example, a new peer with only one good recommendation will have nearly the same reputation value of a peer with m good recommendations.

To mitigate the effect of White Washers, we propose the risk metric given by (1), where r is the number of recommendations that have been emitted about peer X and m the maximum number taken into account in the reputation calculation.

$$Ri_A(X) = \left(1 - \frac{r}{m}\right) \tag{1}$$

The result is a number in the range $[0, 1]$, 0 means no risk, the peer has a sufficient history of recommendations and the reputation value can be taken into account without risk. On the other hand, a risk of 1 means that the reputation value is very risky because there is not enough information about X, and the computed value is the default for new peers.

3.2 Oscillating Personality

The problem of oscillating personality appears because the reputation value is generally an average or a weighted average of the recommendations that have

been emitted about a peer. The result gives a global idea of the past behaviour of the peer.

A peer which makes a good transaction and a bad one in turn will have a reputation value in the middle range, and can be seen like a peer that has an average behaviour. However, this peer is a malicious peer that makes good recommendations to balance its bad behaviour and to continue appearing like an average peer, instead of a malicious one. It can be more interesting for an application to choose a peer with a more regular behaviour than a very irregular one.

We use the standard deviation of the emitted recommendations to detect this kind of behaviour. The bigger is the standard deviation, the farther are the recommendation from the average. A value of 1 means that there is a risk of 100%, and 0 means no risk (i.e. all the recommendations are near to the average value).

The metric in (2) allows to detect an oscillating personality. The role of factor 4 is to normalize the equation to obtain a value in $[0, 1]$ (considering the recommendation values also in this range), r is the number of recommendations used to compute the risk, and $F_i(X)$ is the recommendation of index i about peer X.

$$Ri_B(X) = 4 \times \frac{\sum_{i=0}^{r-1}(F_i(X) - \overline{F(X)})^2}{r} \tag{2}$$

3.3 Random Behaviour

A peer has a random behaviour when the recommendations emitted for this peer are fully distributed in the range of possible recommendations (in our case in the range [0..1]). A Byzantine peer can have this kind of behaviour. From the reputation system point of view, this type of peers will have the same reputation value than ones with a permanent regular behaviour.

This is significantly different from the previous case because for a random behaviour, the standard deviation of the emitted recommendations for this peer will not result in a high value.

Thus, we use the entropy of the recommendations values to detect this type of behaviour. The entropy is an indicator of the level of disorder in the data. A peer with low entropy is a peer with no disorder in the recommendations, which means that its behaviour has always been the same. A peer with a high entropy, is a peer with recommendations values fully dispersed in the range of recommendation.

$$Ri_C(X) = \frac{\sum_{j=1}^{l} p_X(x_j) \log_2(p_X(x_j))}{\log_2(l)} \tag{3}$$

Equation 3 shows the risk metric to detect this kind of behaviour, where l is the number of possible values for a recommendation (cardinality of the set of discrete recommendation values), and $p_X(x_j)$ is the number of recommendations with the value x_j for X divided by the total number of recommendations.

For a reputation system with a continuous range of recommendation values, for example [12] in the range $[0, 1]$, applying this metric requires to make the range discrete. An example of to make a continuous range discrete is that the sub range $[0, 0.2]$ is assigned to $p_X(x_1)$, that is, all the values in that range counts to compute the probability $p_X(x_1)$, the range $]0.2, 0.4]$ is assigned to $p_X(x_2)$, and so on.

The denominator of (3) is a normalization factor. The result is in the range $[0, 1]$. The numerator represents the maximal possible entropy with all the values equally dispersed in the l possible categories of the recommendation values.

3.4 Repeated One Shot Attack

A One Shot Attack occurs when a peer, which is apparently a good one, makes sparse bad transactions. As most of the transactions of the peer are good ones, the bad transactions do not make significant changes to the overall reputation of the peer that will be a good reputation. This is absolutely impossible to detect for an application using only the reputation value.

In the reputation system proposed in [9], a behaviour like the one illustrated in Fig. 1 gives a reputation value of 0.8 (considering equal credibility values for all the evaluators). This value does not show that this peer is a malicious peer which has a malicious behaviour every 3 transactions.

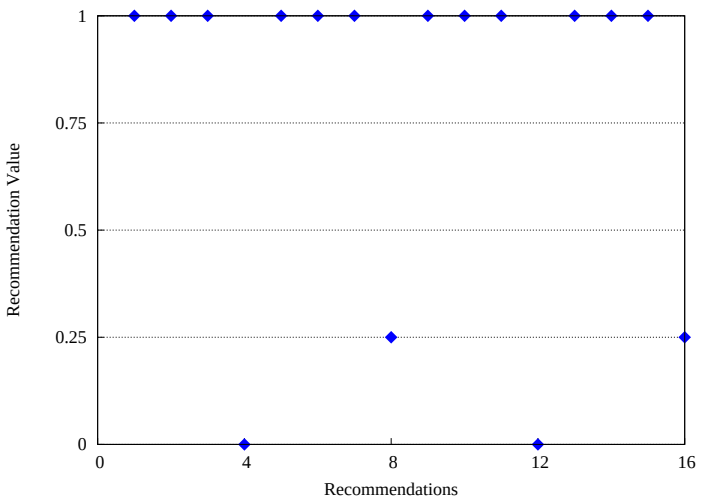


Fig. 1. Example repeated one time attack

The risk metric we propose to detect the Repeated One Shot Attack is based on the analysis of the difference among consecutive recommendation values for a peer. The attack is only possible if the recommendations are clearly partitioned into two groups, with good and bad recommendations (there is no average recommendation), and if there are only sparse bad ones. In this case, the risk metric propose in (4) gives an evaluation of the risk, and 0 otherwise.

Only when there are more stable and good recommendations than bad ones there is a possibility of this attack, for this reason (5) gives 0 risk otherwise. A recommendation value will be considered suspicious if the difference between itself and the previous transaction is bigger than a value D, that depends on the range of the recommendation values. A value of D equal or bigger to $0,5$ would be a adapted difference in a recommendation value range of $[0, 1]$. In (5) r is the number of recommendations the system has about X.

$$J(X,i) = \begin{cases} 1 & if \quad |F_i(X) - F_{i-1}(X)| > D \\ 0 & otherwise \end{cases} \tag{4}$$

$$Ri_D = \begin{cases} \dfrac{\sum_{i=1}^{r} J(i,X)}{r - \sum_{i=1}^{r} J(i,X)} & if \quad \sum_{i=1}^{r} J(i,X) < \frac{r}{2} \\ 0 & otherwise \end{cases} \tag{5}$$

4 Global Metric

We have presented a set of risk metrics to help an application in the decision process to make a transaction with a given peer. A global risk can be computed according to the applications needs. The factors α, β, γ and δ allow the application to give more weight to each term according to its requirement. Equation 6 gives the global risk computation.

$$Ri_{Global}(X) = \frac{\alpha Ri_A(X) + \beta Ri_B(X) + \gamma Ri_C(X) + \delta Ri_D(X)}{\alpha + \beta + \gamma + \delta} \tag{6}$$

The sum of all factors is used to maintain the result within the range [0..1].

To decide whether or not making a transaction with a given peer X an application has two indicators, the reputation $Re(X)$ of peer X, and the global risk value $Ri_{Global}(X)$ associated to X. The use of the reputation value and the risk value completely depends on the application needs. For example, for applications that need a higher level of security they can have a boundary of the risk value to make a transaction, and for applications that need a lower level of security these two values can be aggregate with a weighted average function.

The reputation value of a peer with a low risk means that the reputation value effectively reflects the past behaviour of the peer. A high risk means that

the reputation value does not necessarily reflects the past behaviour of the peer, and making a transaction with this peer may be hazardous. Nevertheless, a high risk does not means that the peer is a malicious one, it is only a high probability, and the transaction may succeed.

It is worth to mention that two other types of malicious behaviours that were not considered in this work: milking personality and false recommendations. The reason is because they can be easily detected during the reputation value calculation.

Milking personality is the strategy of a peer that builds a good reputation value and after some time starts having a bad behaviour. As its reputation value is high, the peer can deceive other peers until its reputation value falls. To detect this behaviour the metric for the reputation value can add a fading factor, which gives more weight to the latest recommendations. False recommendations are the recommendations emitted by malicious peers about other peers, but they do not reflect the peer's behaviour during the transaction. The system can use a credibility value to detect this behaviour.

In the next section, we present some simulation results to show the efficiency of our metrics.

5 Results and Analysis

The experiments have been done in order to quantify the efficiency of the risk metrics in front of the correspondent attack. All of them have been done using the reputation system proposed in [9]. This reputation system uses a list of the last m recommendations emitted about a peer to compute its reputation value. In the experiments the size of the recommendation list has been set to $m = 16$, because this value has shown to be the best choice for this reputation system (See [9]).

In all the experiments the total number of peers simulated is $100,000$, which perform an average of 100 transactions each. The results were obtained averaging the results every $200,000$ transactions. For each transaction, a peer A randomly chooses a peer B in the network to make the transaction. To decide whether or not to make the transaction the risk and reputation value are aggregated using (7). This value is used as a threshold to probabilistically decide to accept or deny the transaction. The key idea in (7) is to increase or decrease the threshold according to the reputation and risk values. Increasing the threshold for peers with bad reputation has the purpose of not allow the inanition of transaction for some peers in front of the attack of false recommendations, and because a high risk for a low reputation value means that the reputation does not reflect the real behaviour of the peer.

$$Th_t(B) = \begin{cases} \text{If} & 0.75 < Re_t(B) \leq 1 & Re_t(B) \times \left(1 - \frac{Ri_t(B)}{2}\right) \\ \text{If} & 0.25 \leq Re_t(B) \leq 0.75 & Re_t(B) \times (1 - Ri_t(B)) \\ \text{If} & Re_t(B) \leq 0.25 & Re_t(B) \times (1 + 2 \times Ri_t(B)) \end{cases} \quad (7)$$

The first experiment is about White Washers. 20% of peers in the system are White Washers. They make malicious transactions and when their reputation value drops down to 0.05 they leave the system and join again with a clean new identity.

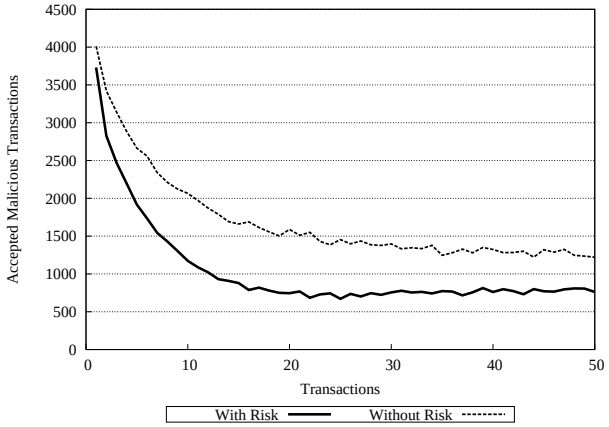


Fig. 2. Accepted Transactions to White Washers

Figure 2 shows the accepted transactions to white washers in the reputation system with the risk metric and without it. The axis X in the figure represent the result obtained every 200.000 transaction, from a total of 10.000.000 (50 results). Axis Y represents the number of accepted transaction in each set of transactions. As we can see in Figure 2 malicious transactions decrease in more than a 40%.

In this case, the risk metric affects the new honest peers in the system, but as they continue to do honest transactions to obtain good recommendations, the risk value rapidly falls to 0 and stops affecting the transactions between these peers. Figure 3 shows the evolution of the risk value for honest peers and for the malicious ones. The results represents the average of the risk of the set of peers. We see in this figure that the risk for the honest peers goes down as they make more transactions in the system.

The second type of behaviour to analyze is the oscillating personality. In this experiment we have considered malicious peers that make a good and a bad transaction in turn to continue with a regular reputation value. The results are showed in Fig. 4. The accepted malicious transaction drop in more than 80%, which shows that our metric is very efficient to detect this kind of behaviour. In this case, honest peers are minimally affected by the risk metric since they usually make good recommendations. Moreover, the number of false recommendations is not sufficient to get a high risk.

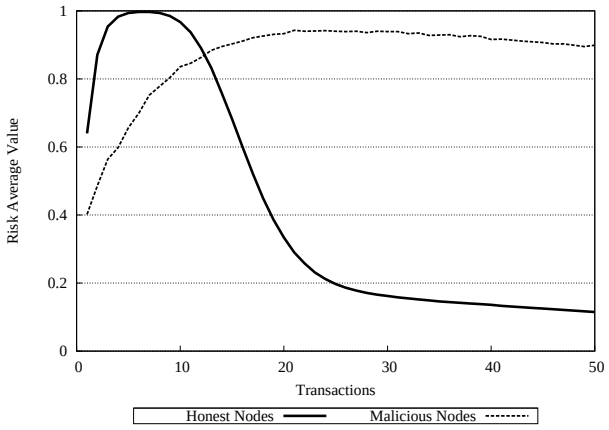


Fig. 3. Risk Value evolution with White Washers

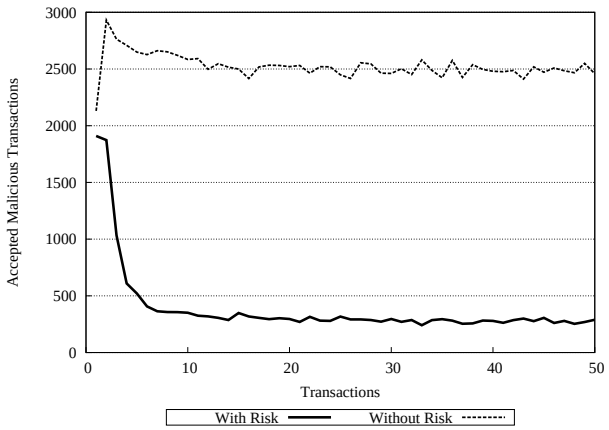


Fig. 4. Accepted Transactions to Oscillating Personality

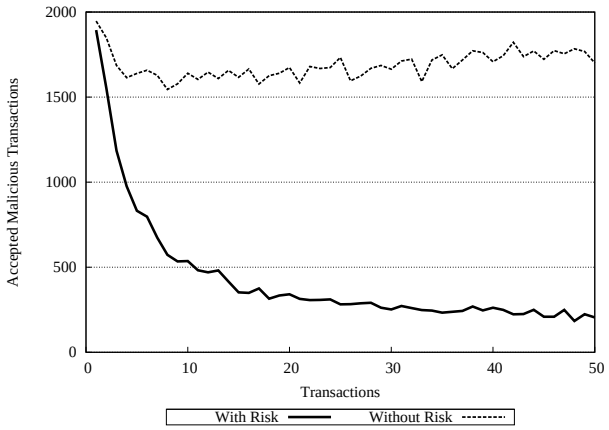


Fig. 5. Accepted Transactions to Random Behaviour

The results for the analysis of the metric presented for the random behaviour are presented in Fig. 5. This figure shows that without the risk metric, 20% of the malicious peers make 1500 bad transactions. Using the risk metric based on the entropy, the number of malicious transactions falls under 250, which represents an improvement of more than 80%.

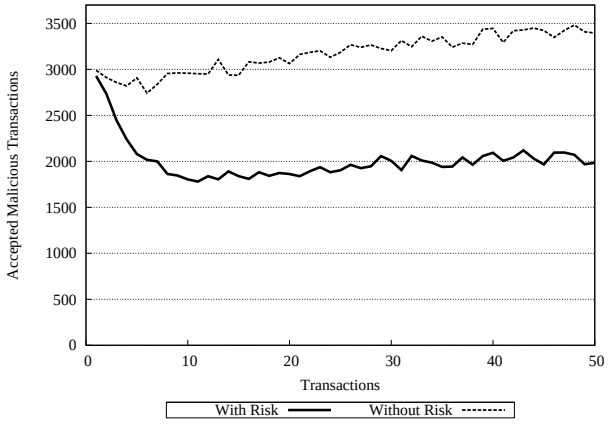


Fig. 6. Accepted Transactions to Repeated One Time Attack

The last experiments analyze the behaviour of the metric in front of the Repeated One Shot Attack. The parameter D have been set to 0.5 which is half of the total range. In this case, we have considered malicious peers that

repeatedly make 3 good transactions and then a bad one. The results are shown in Fig. 6.

This metric avoids making around 40% of malicious transactions. Honest peers are only affected by this metric if there are false recommendations in the system. If there is a high percentage of lying peers, the metric could think this is a Repeated One Shot Attack. This really depends on how long is the list of recommendations considered in the risk and reputation computation

6 Conclusion

This works introduces the concept of risk metric in reputation systems to complement the reputation value and to detect some suspicious behaviour ignored by the reputation value. We have presented four risk metrics based on the analysis of the list of recommendation the reputation system has about a given peer.

The experiments have shown very good results in the detection of the attacks and a clear fall in the number of malicious transaction made by peers with wrong behaviour (up to an 80%). The risk that has been proposed helps to trust the reputation value itself, preventing an application from making very hazardous transactions.

Further efforts have to be made to detect other kinds of attacks to reputations systems. We especially think about the detection of White Washers which is a difficult task for reputation systems.

Further work also consists in creating risk metrics for other models of reputation systems, like the ones based on transitive reputation. Another pending issue is to test different aggregation schemes for the risk and the reputation value, depending on the requirements of the applications.

References

1. Xavier Bonnaire and Erika Rosas. A critical analysis of latest advances in building trusted p2p networks using reputation systems. In *WISE2007 Workshops: The 8th International Conference on Web Information Systems Engineering*, Lecture Notes in Computer Science 4832, pages 130–141. Springer-Verlag, December 2007. (To appear).
2. Ernesto Damiani, Sabrina De Capitani di Vimercati, Stefano Paraboschi, and Pierangela Samarati. Managing and sharing servents' reputations in p2p systems. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):840–854, July/August 2003.
3. John Douceur. The sybil attack. In *IPTPS '02: Proceedings of the First International Workshop on Peer-to-Peer Systems*, volume 2429 of *Lecture Notes in Computer Science*, pages 251–260, Cambridge, MA, USA, March 2002. Springer.
4. Eric Friedman and Paul Resnick. The social cost of cheap pseudonyms. *Journal of Economics and Management Strategy*, 10(2):173–199, June 2001.
5. Anurag Garg, Anurag Garg, Roberto Battiti, Roberto Battiti, Gianni Costanzi, and Gianni Costanzi. Dynamic self-management of autonomic systems: The reputation, quality and credibility (rqc) scheme. In *In The 1st IFIP TC6 WG6.6 International Workshop on Autonomic Communication (WAC*. Springer-Verlag, 2004.

6. Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*, pages 640–651, New York, NY, USA, 2003. ACM Press.
7. So Y. Lee, O-Hoon Kwon, Jong Kim, and Sung J. Hong. A reputation management system in structured peer-to-peer networks. In *WETICE '05: Proceedings of the 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise*, pages 362–367, Washington, DC, USA, June 2005. IEEE Computer Society.
8. Zhengqiang Liang and Weisong Shi. Pet: A personalized trust model with reputation and risk evaluation for p2p resource sharing. In *HICSS '05: Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, page 201.2, Washington, DC, USA, January 2005. IEEE Computer Society.
9. Erika Rosas. Diseño e implementación de un sistema de reputation para redes p2p mixtas. Master's thesis, Universidad Técnica Federico Santa María, November 2007.
10. Aameek Singh and Ling Liu. Trustme: Anonymous management of trust relationships in decentralized p2p. In *P2P '03: Proceedings of the IEEE International Conference on Peer-to-Peer Computing*, page 142, Washington, DC, USA, September 2003. IEEE Computer Society.
11. Mudhakar Srivatsa, Li Xiong, and Ling Liu. Trustguard: countering vulnerabilities in reputation management for decentralized overlay networks. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 422–431, New York, NY, USA, 2005. ACM Press.
12. Li Xiong and Ling Liu. Peertrust: supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Transactions on Knowledge and Data Engineering*, 16(7):843–857, 2004.
13. Bin Yu, Munindar Singh, and Katia Sycara. Developing trust in large-scale peer-to-peer systems. In *MAS&S2004: Proceedings og the IEEE First Symposium on Multi-Agent Security and Survivability*, pages 1–10, Philadelphia, Pennsylvania, USA, August 2004. IEEE.
14. Runfang Zhou and Fellow-Kai Hwang. Powertrust: A robust and scalable reputation system for trusted peer-to-peer computing. *IEEE Transactions on Parallel and Distributed Systems*, 18(4):460–473, April 2007.

Extending the Comparison Efficiency of the ART Testbed

Yann Krupa, Jomi Fred Hubner, Laurent Vercouter

École des Mines de Saint-Étienne
Centre G2I, Équipe SMA
158 cours Fauriel, 42023 Saint-Étienne Cedex 02, France
{krupa,hubner,vercouter}@emse.fr

Abstract. In online communities, systems use reputation and trust values to make people more comfortable in doing business with unknown partners. A lot of research about trust and reputation models has been done in the fields of psychology, sociology, and more recently in multi-agent systems. Many models have been proposed to decide either to trust, or not, some other agent in a given context. In this article, we analyse the benchmark currently used for trust models comparison, the ART testbed. Based on critical feedbacks given by ART users and coming from our own experience, we emphasize its limitations. We suggest a new approach using several scenarios to extend the comparison efficiency of the testbed. Two complementary scenarios are also proposed as an illustrative example of this approach.
Keywords: reputation, ART, trust, agent, testbed.

1 Introduction

Trust can be defined as a mental state that is reached when both the truster expects that the trustee will behave in a given manner, and when the truster accepts the risks related to the failure of the interaction [4]. While distributed systems are rising, trust is making its way towards online applications as privacy and security cannot be maintained by conventional means. A lot of research has been done in the multi-agent field to provide trust and reputation models, in many application domains.

To compare those different models and to provide an experimental standard, the ART testbed [7, 8, 1] has been defined. The testbed simulates an art appraisal application where appraisers rely on others to evaluate art items. Three years after its creation, ART is recognised in the community and is now used as a reference by researchers. Nevertheless, ART has some drawbacks we underline in this paper. Some articles already discussed some of ART's drawbacks and the improvements that can be made over the testbed [10, 19, 18]. Finally, it seems hard to implement real models on ART due to its particularity and the low number of available information sources.

We propose here a complementary approach, by using multiple scenarios instead of a single one. First, we define a method for evaluating trust scenarios upon the expressiveness that each scenario allows for the models. Then, new

scenarios are defined as an illustrative exemple of how the comparison efficiency of ART could be extended by using our approach.

In the first part of the article, we review the ART testbed and underline its main problems. Next, we propose a method to evaluate trust scenarios. The new scenarios are then defined as an exemple of our approach.

2 The ART Testbed

Due to the heterogeneity of their application domains and specificities, trust models are difficult to compare. Each author has its own way of evaluating his model. The ART workgroup was created in order to provide a comparison standard for trust and reputation models, allowing evaluation and experimentation [9]. It is used in a competition in order to compare the existing approaches.

2.1 ART's Art Appraisal Scenario

On ART, each "participant" must provide an agent implementing a trust model. This agent takes the role of an art appraiser who gives appraisals on paintings presented by its clients. To fulfill his appraisals, the agent asks opinions to other appraisers. These agents are also concurrents and free of their actions and thus, they may lie in order to fool opponents.

The testbed provides a "simulator" that supervises the game, handles the clients, and so on. The simulation runs in a synchronous and step by step manner. The scenario evolved during the years, the 2008 version is the one explained here, a detailed explanation can be found on the website [1].

Each simulation step goes like this:

- Clients (handled by the simulator) ask appraisers for opinions on paintings . Each painting belongs to an era. Appraisers are the agents implemented by the participants.
- Each appraiser has a specific expertise level in each era. The error an appraiser makes while appraising a painting directly depends on this value and the money the appraiser decides to spend for that appraisal.
- An appraiser cannot appraise its paintings himself, he must ask other appraisers for appraisals, thus pushing the appraiser towards a situation in which he has to rely on others.
- As agents are allowed to lie, each one should maintain a trust model in order to anticipate others behavior. Agents can purchase opinions about an agent to other players (they can lie, *i.e.* tell that someone honest is a liar and vice versa), this is called the "reputation protocol".
- Agents weight each received appraisal in order to calculate the final evaluation.
- The accuracy of appraiser's final evaluations is compared to each other, thus determining the client share for each appraiser during the next turn (the most accurate receive more clients). At each turn, an appraiser earns money from his clients and spend some asking others advice.

– When the turn ends, the simulator reveals the real value of each painting. Agents can then spot liars or begin to be suspicious towards some agents that may have lied.
The winner is the appraiser agent with the highest bank account at the end of the game.

2.2 Scenario's Limits

In this section we list some of ART's drawbacks that have been revealed either by participants of ART competitions, our own use of the testbed or previous ART analysis [10].

Reputation Issues One of the first problem that was underlined after the first ART competition was the uselessness of reputation protocol. Winners of the 2006 competition [18] underlined 2 facts about it:

1. Reputation semantic is hidden and ambiguous. It's a simple real value between [0,1] mixing different criteria, including among others skill and honesty. So if agent X tells W that Y has a reputation of 0.13, W will not know if Y is a liar, a bad appraiser, or a bad reputation provider.
2. The number of players in the game is really low, it is easy to learn their behavior. After a few turns it is possible to tell who lies, and who doesn't.

Eventually, the IAM team decided not to implement the reputation protocol at all. Reputation is second hand information (because transmitted by other agents), so it is less reliable than direct interaction information. A model will probably use reputation only in cases where direct information is lacking. On ART, every agent has around 20 paintings to appraise per step, each one requires 1 or 2 advices, giving a (mean) total from 20 to 40 direct interactions per step. If the number of agents in the competition is 10, each agent will interact directly 2 to 4 times per step with each other. Considering this, it does not seem necessary to use the reputation protocol. The agent that won the 2008 contest, UNO, doesn't use the reputation protocol either, underlining that the asked agent may not have sufficient knowledge about whom is asked [13].

Trust Model Simplification While implementing an agent for the AAMAS'08 competition, we faced some difficulties that raised our interest: one of our objectives was to implement the LIAR model [12], which is a model dedicated to P2P networks. It uses a lot of information sources, and some communication specificities. As ART only provides direct information and reputation messages, we implemented a dramatically simplified version of LIAR. We eventually ended up with a model really different from LIAR. This is problematic when using ART because the goal of the testbed is to evaluate trust models, but eventually, due to the huge simplifications, we can't say that it is the LIAR model that has been evaluated.

Parameter Tuning One of the hardest point while setting up our model was the parameter tuning. Liar detection and weight providing requires a deep understanding of the scenario (more precisely of the appraisal calculation function) in order to be well tuned. From our point of view, these difficulties are out of scope for a trust model.

Honesty or Cheating? What should the agent answer when asked for opinion or reputation? will it lie or not? Does this choice has an outcome over the contest result? What happens if everybody decides to provide a "full-time lying agent"? or at the contrary, agents that never lies? Table 1 shows an experiment done with some of the 2008 contest participants. Our agent, called Simplet, has been splitted in 2 versions, one that is always honest, and one that always lies[1]. The experiment has been done on 5 runs on ART, the score column represents the total amount of money won over those runs. On the first series, Honest Simplet was 325 000 behind the leader, whereas on the second one, Lying Simplet is only 35 000 behind. Moreover, he goes from the 5th to the 3rd position. We can see here that there can be a clear difference between the two outcomes depending on which strategy is used. This is a problem as ART is willing to measure trust model's performance, and as shown here, not only the trust model is evaluated: without changing the model, the results changed significantly. Note that we can explain that some other models (*e.g.* FordPrefect) changed positions because of their sensitivity to Simplet lies.

Honest Simplet		Lying Simplet	
Agent	Score	Agent	Score
Uno2008	1 351 850	Uno2008	1 251 992
FordPrefect	1 307 270	connected	1 246 134
connected	1 179 149	**Simplet**	1 217 371
Next	1 109 100	FordPrefect	1 181 958
Simplet	1 027 946	Next	989 904
IAM	666 080	ArtGente	718 957
ArtGente	659 293	IAM	621 277
MrRoboto	519 970	Peles	577 343
Peles	502 174	MrRoboto	504 251

Table 1. Experiment on ART, same trust model, different answering strategy.

[1] Honest Simplet (trustworthy) always answers as good as he can when asked for opinion about a painting by some other appraiser. Lying Simplet (untrustworthy) simply returns an erroneous appraisal when asked for it.

Open Systems Multi-agent systems are meant to be open: this is not ART's case. Agents join the game at the beginning, and quit after the last turn. Nobody leaves or enters during the game. System's openness in trust scenarios brings out new and complex situation, it is therefore interesting to allow them. Openness often raises problems, when a new user joins a system, he often has a "zero reputation". People tend to be really suspicious towards newcomers. It is hard to decide how to handle unknown agents, either you take some risk by interacting with them, or you decide not to interact and you may end up alone.

ART's drawbacks have been explained here in a descriptive way. We need to define a method for comparing trust scenario's drawbacks and advantages. This is what is done in the next section.

3 Means for Scenario Analysis

In order to compare trust scenarios and to provide a clear view of each scenario's drawback and advantages, we define here a method for scenario analysis. This method evaluates each scenario based on the expressiveness it allows for the trust models. Our approach is based on the following statement: If a model uses a given criterium in order to take its trust decision, then if a scenario does not provide this source, the model evaluation will be biased. We list here the main criteria that are present in the domain, thus allowing to define multiple scenarios providing those criteria. A good coverage of the research domain can then be achieved by a set of scenario.

The criteria that follows have been inspired by J. Sabater's state of the art [16].

3.1 Criteria

We list here the criteria used to evaluate trust scenarios and give a short explanation for each one of them.

There are two main groups of criteria, the first one is the "Information sources". These are informations regarding the other's agents behavior.

- DI: Direct Interaction. This is the basic information source, when agent X interacts with agent Y, then X and Y can both get an idea about the other's behavior.
- DO: Direct Observation. An agent Z can observe an interaction between X and Y. This information is less frequent in real world applications , you can find it in some networks where you can "hear" things without interacting (overhearing).
- WA: Witnessed Appreciation. Z tells Y about what he thinks of X. This is what is usually designed by "gossip" .
- WF: Witnessed Fact. Z tells Y about what X did. This let Y judge by himself what he is told about [12], e.g.: Z is using proprietary software, if X is an open source advocate he will judge Z action as a bad action. On the contrary, if X works for a software company, he will be pleased.

- SI: Sociological Information. Agent X can infer Z reputation by knowing some sociological information, e.g.: X knows that Z works with Y, whom X trusts a lot, he can then infer Z reputation by saying "there are chances Y will not work with an untrustworthy agent".
- P: Prejudice. X can judge Y just by observing his characteristics. This is the default judgement: without interacting, we use all the information we have at our disposal to judge an agent. *e.g.* A delivery boy knocking at our's door in a uniform will be easily trusted whereas the same boy without his uniform will not [6].

The second group of criteria concerns the interaction context and the general environment specificities.

- Reputation visibility [16]. In a system like eBay, reputation visibility is global, it means that anyone can see all the information concerning the reputation of an other agent. On the opposite, on ART for example, reputation visibility is subjective. It means that for an agent to know an other agent reputation, he will have to ask others about it.
- Multi-context Granularity. Does the scenario provides multi-contextual granularity? The trust value associated to an agent will depend on the context: *If we trust a doctor when she's recommending a medecine it does not mean that we have to trust her when she is suggesting a bottle of wine* [16].
- Test interactions. Does the scenario allow low cost, low risk interactions? Repage model [14] uses low risk interaction when the agent is unable to decide whether to trust or not. If I'm willing to buy a rare and expensive collection stamp from someone I can't decide if he's trustworthy or not, I'd buy a far less expensive stamp just to get a better idea of this seller trustworthiness.
- Warranties. Is it possible to purchase warranties, to sign contracts or to ask for third party services? Contracts, promises, warranties and third parties services are underlined by C. Castelfranchi and R. Falcone [5] as they can increase the risk acceptance level.
- Stake. Are risks, utility, and importance different from one interaction to another? For example, buying a pen to an unknown seller is less risky than buying a car to the same person (less funds are at stake). Importance and utility were already used in one of the first models, to take the trust or distrust decision [11]. We can illustrate the stake in a different example: if somebody is a stamp collector who has been looking for a particular stamp for a long time and finally finds it, owned by a seller who he is not sure about, this seller will buy it, accepting the risks due to the importance and utility of the outcome.
- Openness. Is the system open? Can agents join (or leave) the scenario during the game process? In many applications for the trust problems, the system is opened. This means that an agent can leave or join whenever he wants. This is a big problem in trust: the model must be suspicious towards new comers, but not xenophobic.

– Homogeneity. Does the scenario allow games to be played with all agents having the same model? Some models will probably work better if the other agents use the same model, for example, some may require that every agent handles a trustnet [17]. This consideration may be interesting in some specific fields of application, like a P2P network where all peers use the same reputation model, allowing cheaters (modified clients) to be easily spotted. For a scenario to allow homogeneity, it must allow agents using the same model to play versus the testbed. The final score (sum of all model scores) will then represent the model adaptive power towards the scenario rather than towards the other models[2].

3.2 Evaluating ART's Scenario

The criteria that have been defined in the previous section are summarized in a grid. They are applied to the ART scenario to evaluate its expressiveness and domain coverage. Results are shown in Table 2. The main drawbacks of ART are also summarized by the grid.

The grid will be filled this way:

– Y: criterium fulfilled by this scenario,
– empty cell: unfulfilled criterium,
– S: Subjective visibility (Vis criterium),
– G: Global visibility (Vis criterium),
– M: Multi-context granularity (Gran criterium),
– Si: Single-context granularity (Gran criterium).

	Information sources						Environment specificities						
Game	DI	DO	WA	WF	SI	P	Vis	Gran	Test	Warr	Stake	Open	Homo
ART	Y		(Y)				S	M					

Table 2. Criteria grid, applied to ART.

There's a large amount of direct interactions (DI) in ART, but it is not possible for another agent to observe those. Reputation (under WA form) exists in ART but is quite unused by participants, whereas there is no sociological information and the scenario does not provide any means of using prejudices.

Regarding the environment settings, reputation has a subjective visibility as each agent must ask others to receive reputation messages. ART provides a multi-context granularity on the eras, as for each era, an agent may be trusted

[2] Thus, zero-sum games do not allow homogeneity as the sum of all agent scores will always be equal to zero. This is also true for ART where the sum is equal to the number of clients multiplied by the number of game steps.

differently. All the opinion requests have the same cost on ART, it isn't possible to do low cost, low risk test interactions, neither to use stake appreciation to take the trust decision. But it is possible to do normal cost, low risk interaction, when not sure about a given agent: ask for opinion and then provide a zero weight. This will allow to check afterwards if the agent could have been trusted or not without exposing ourselves to its potential lies.

ART scenario does not provide any kind of warranty. The scenario is closed, no agents can enter or leave during the game. Finally, ART is not defined to allow homogeneity.

4 Extending the Scenario's Set

ART comes from a great challenge: regroup all the trust actors under a single standard scenario. But this goal is hard to achieve, the application domains can be really different from one model to another and we do not think there is an ultimate trust scenario that can regroup all the aspects involved in trust. We propose a solution between the "pre-ART" situation, which leads to one scenario per model, and ART, which leads to one scenario for all models. The solution is a proposal of a set of scenarios covering different aspects of trust and that can therefore be associated with different applications. Thus, a model can be implemented on one, some or all the scenarios of the competition. Then someone with an applicative problem should just look at the scenario (or the criteria) which is the closest to his application to find the most relevant model for this problem.

In this section we propose two new scenarios as an example of how a good coverage of the trust domain can be obtained by using a set of scenarios. They are complementary to ART in the fulfilment of the criteria enumerated in the previous section. The grid allows to evaluate quickly the domain coverage of the different scenarios and of the set. We also want the scenarios to allow the evaluation of the trust models separately from the agent himself.

4.1 Trust Game

This game was used by economists [2] to check the "Homo œconomicus model", upon which an economic man will prefer to keep the money he has instead of risking to lose some.

The original game is the following:

- 2 players (P1,P2), who cannot communicate and don't know each other are put in separated rooms,
- the organizer gives 4$ to P1 and P2,
- P1 can then decide to give 0,1,2,3 or 4$ to P2, knowing that the researcher will triple it before giving it to P2,
- P2 receives the money P1 sent multiplied by 3, he then decides how many he wishes to send back to P1 (from 0 to everything).

– both players leaves.

This game is in fact a generalization of trust problems in which someone decides whether to trust someone else or not, and with what level of involvement. A greater involvement increases both the loss and gain possibilities.

In the original version both player leave after the game, because the economists do not want the fear from reciprocity to intervene. If the game was iterated, P2 could fear that the next time he will encounter P1, this last one would not be generous if P2 have not been before. This would have changed the experiment.

In our case, our objective is slightly different and an iterated version of this game is interesting in order to spread reputation. Each agent knows who interacted with who, and can then ask for reputation between the iterations. The idea of this game is to work on other reputation sources that direct interaction, in order to encourage the use of reputation. In online markets and in many situations, direct interactions are quite rare between two given agents. In that scenario, we propose an extreme solution: each couple of agents will only interact once in the game. Doing so, agents will be forced to rely on others to determine whether it's a good idea to trust or not. In our version, the multiplier (originally set to 3) is variable, thus introducing stake. It is worthier taking the risk of interacting when the multiplier is high. Artificial prejudices are defined by creating agent groups based on their strategy. For example, the game could create a group 1 with 80% of generous agents, an other group 2 with 60% of non generous agents... While interacting with a given agent, it will then be possible to know from which group this agent is (but it would not be possible to know how the game created the groups). Thus, the model could associate a trust value to a certain characteristic (which would be the group number).

The number of agents in the game should be high (at least above 50) to make it interesting. In order to resolve this problem along with the problem of evaluating the model separately from the agent strategy, we propose an agent "architecture" for this scenario. On one side of the agent, the model will implement all the trust and reputation functionalities in an honest way (no lies): it will decide whom to ask for reputation, how many to send to P2 and build agent reputation. On the other side, the strategic module will implement honest or dishonest functionalities regarding the scenario's strategy: it will compute how many dollars to return to P1 after he sent this agent a given amount of money, to modify or not a reputation message emitted by the model (in order to lie), ...

This architecture allows these things:

– a large number of agents can be made by combination of different models and strategies,
– model, agent and strategy can be evaluated separately, given that all money earned by playing P2 role is kept in a specific bank account for the strategy, and the money earned while playing P1 role is kept in the model's account.
– it is then possible to run the game with all agents having the same model (homogeneity criterium).

4.2 Online Market

Our second scenario is inspired by online markets, our goal here is to get a bit closer from online applications of trust.

The participating agents in this game are buyers. They are given a list of items to purchase and a budget.

Sellers (potentially untrustworthy ones) are controlled by the simulator, they put items on sold for a given time (step number) and a fixed price. For example, Seller X is told to put bikes on sale during 3 steps at 500$ per bike. For equality reasons between participants, sellers have unlimited stock. Whereas the limited time during which a given seller proposes a given object leads to situations in which the buyer will be urged to take a decision whether to purchase or not.

This can lead to a situation in which only an untrusted seller provides the item. In that case, a good trust model will either:

- engage in a low cost low risk interaction, if the provider is selling low cost objects along the required item,
- purchase a warranty: by paying 10% of the item cost to the sim, this last one will refund to 60% if the seller decides not to send the item after receiving the money.
- engage a third party: the buyer can ask a trusted seller to take the third party role by paying a constant price. The third party will then receive the money from the buyer (item price and honoraries), he will contact the seller and ask for the object. If the seller refuses to send the object, the buyer will be completely refunded the item price.

Buyer communicate using WA and WF between turns. As it is inspired by online communities, reputation is global (each agent carries all the advices concerning him), this allows the possibility of doing experiences with results that can be exploited by online markets.

The game ends after a known number of time steps. The game itself is iterated (without resetting agent memories) a certain number of times to prevent border effects. The winner is the agent with the maximum amount of object (each object has a value equivalent to its price).

Finally, this game is not required to be played with a large amount of agents, but it is designed to be open: during the game, sellers will left and others will enter the game, introducing the openness problem.

4.3 Synthesis

We can use the analysis grid (cf. Section 3) to get a general view of the interest of the scenarios, the results are presented in Table 3. Direct Observations could be added quite easily to any of the scenarios but will not have a real interest excepted in a specific scenario close to an application in which DO are important.

Sociological Information seems hard to simulate, therefore special scenarios for social aspects should be made from real data like the one coming from social networks. It would have been possible to add a basic sociological information like

in the first version of Regret [15] where an agent can inherit its group reputation. But in fact this is not rich SI, this is more a Prejudice based on the group.

As new scenarios and criteria are available, the models are less restricted and the testbed comparison efficiency is improved.

The table shows how our approach (with the analysis grid) can be used to evaluate trust scenarios, and the coverage of the research domain they provide. Direct Obervations and Sociological Informations are missing, but our first objective here is the approach, not the scenarios themselves. Nevertheless, a good coverage is achieved by the set of scenarios (ART and the 2 example scenarios), as there is almost one "X" in each column.

	Information sources						Environment settings						
Game	DI	DO	WA	WF	SI	P	Vis	Gran	Test	Warr	Stake	Open	Homo
ART	Y		(Y)				S	M					
Trust Game	(Y)		Y	Y		Y	S	M					Y
Online Market	Y		Y	Y			G	M	Y	Y	Y	Y	Y

Table 3. Criteria grid, applied to ART and the two new scenarios.

The set of scenarios solves some of ART's problems listed in section 2.2:

Reputation Issues

- Reputation protocol is said useless on ART: On Trust Game, the number of DI is so reduced that models can only rely on reputation. In the Online Market scenario, sellers enter and leave during the game, thus the number of DI between two given agents will be low. Moreover, the Global Visibility of reputation makes it easier to access.
- The second reputation problem that is addressed is about the low number of agents, making easy to learn who lies and who does not. We propose a scenario working with a large number of agents, and a second one allowing openness. Both solves the problem of learning opponent's strategies.

Trust Model Simplification Since there are more Information sources and Environment settings available, models will need less simplification while implemented on the testbed. Nevertheless, there is still a need for simplification and adjustment as models are not defined specially for a given scenario. The only solution is either to define a scenario specially for a given model, or to design a model specially for a scenario.

Parameter Tuning On ART it is hard to detect a lie because it needs a deep understanding of the appraisal calculation function. On the new scenarios, there

is no hidden mechanism, no black box and no complex functions. Parameter tuning will be easier as we have perfect knowledge of the game.

Honesty or Cheating? We proposed in both scenarios a separation of model and strategy components. In the first scenario, an agent is composed of two, independant but communicating, parts: model (trust or don't trust) and strategy (lie or don't lie). In the second scenario, there are two kinds of agents: buyers (implementing the model) and sellers (implementing the strategy). A problem that has not been solved is to know who should implement the strategy? It could be the organizers of the ART contest, but in this case we take the risk of defining a strategy set too restricted. Otherwise, the participants can implement these, but in that case we take the risk of having agents defined specially to be compliant with the model of that participant.

Open Systems The Online Market scenario allows openness.

Our goal here is not to show how that new scenarios are perfect, because they are not! Moreover, the scenarios are only given as examples. The point here, is to see how **a set of scenarios** can solve the problems we were facing.

While defining new scenarios, one should keep in mind that a scenario must reflect real life problems and avoid toy problems.

5 Conclusion

Although ART has been contributing as a common testbed for trust and reputation models, it has some drawbacks. We listed them in this article and proposed a solution, along with a new approach for the definition and evaluation of scenarios. The lack of reputation has been solved by the Trust Game scenario which has very few direct interactions, thus making the agents rely on other sources. The problem of reputation's semantic has already been handled with a specific ontology for reputation [3].

Implementing a real model as an agent on a game is still not easy, but now, instead of trying to force it into ART, it's possible to find the scenario which is the closest from the model and make it fit onto it.

Another question that was raised concerned the evaluation of the model that is noisy under ART, because the agent in its whole is evaluated. Both scenarios we proposed suggest separation between the model and the other strategic or lying concerns.

Finally, the ideas proposed in this article will be submitted to the ART workgroup for discussion.

References

1. ART. The art testbed. `http://www.art-testbed.net/`.
2. J. Berg, J. Dickhaut, and K. McCabe. Trust, Reciprocity, and Social History. *Games and Economic Behavior*, 10(1):122–142, 1995.
3. S. Casare and J. Sichman. Towards a functional ontology of reputation. *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 505–511, 2005.
4. C. Castelfranchi and R. Falcone. The dynamics of trust: From beliefs to action. *Proceedings of the Second Workshop on Deception, Fraud and Trust in Agent Societies*, pages 41–54, 1999.
5. C. Castelfranchi and R. Falcone. Social trust: a cognitive approach. *Trust and deception in virtual societies table of contents*, pages 55–90, 2001.
6. B. Esfandiari and S. Chandrasekharan. On how agents make friends: Mechanisms for trust acquisition, 2001.
7. K. K. Fullam, T. B. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. S. Barber, J. S. Rosenschein, L. Vercouter, and M. Voss. A specification of the agent reputation and trust (art) testbed: experimentation and competition for trust in agent societies. In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 512–518, New York, NY, USA, 2005. ACM.
8. K. K. Fullam, T. B. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. S. Barber, J. S. Rosenschein, L. Vercouter, and M. Voss. The Agent Reputation and Trust (ART) Testbed Game Description (version 2.0). 2006.
9. K. K. Fullam, T. B. Klos, G. Muller, J. Sabater-Mir, Z. Topol, K. S. Barber, J. S. Rosenschein, and L. Vercouter. The Agent Reputation and Trust (ART) testbed architecture. In 8^{th} *Workshop on Trust in Agent Societies*, 2005.
10. M. Gomez, J. Sabater-Mir, J. Carbo, and G. Muller. Improving the ART-Testbed, Thoughts and Reflections. *Proceedings of the Workshop on Competitive Agents in the Agent Reputation and Trust Testbed at CAEPIA-2007, Salamanca, Spain*, pages 1–15, 2007.
11. S. Marsh. *Formalising Trust as a Computational Concept.* PhD thesis, 1994.
12. G. Muller and L. Vercouter. L.i.a.r.: Achieving social control in open and decentralised multi-agent systems. Technical Report 2008-700-001, ENSM-SE / G2I, 2008.
13. J. Murillo and V. Muoz. Agent UNO: Winner in the 2007 Spanish ART Testbed competition. *Proceedings of the Workshop on Competitive Agents in the Agent Reputation and Trust Testbed at CAEPIA-2007, Salamanca, Spain*, 2007.
14. J. Sabater, M. Paolucci, and R. Conte. Repage: REPutation and ImAGE Among Limited Autonomous Partners. *Journal of Artificial Societies and Social Simulation*, 9(2):3, 2006.
15. J. Sabater and C. Sierra. REGRET: reputation in gregarious societies. *Proceedings of the fifth international conference on Autonomous agents*, pages 194–195, 2001.
16. J. Sabater and C. Sierra. Review on Computational Trust and Reputation Models. *Artificial Intelligence Review*, 24(1):33–60, 2005.
17. M. Schillo, P. Funk, and M. Rovatsos. Using Trust For Detecting Deceitful Agents In Artificial Societies. *Applied Artificial Intelligence*, 14(8):825–848, 2000.
18. W. Teacy, T. Huynh, R. Dash, N. Jennings, M. Luck, and J. Patel. The ART of IAM: The Winning Strategy for the 2006 Competition. In *Proceedings of The 10th International Workshop on Trust in Agent Societies*, pages 102–111, 2007.

19. L. Vercouter, S. J. Casare, J. S. Sichman, and A. Brandão. An experience on reputation models interoperability based on a functional ontology. In *IJCAI*, pages 617–622, 2007.

This article was processed using the LaTeX macro package with LLNCS style

On the Effects of Reputation in the Internet of Services

Stefan König[1], Tina Balke[1], Walter Quattrociocchi[2], Mario Paolucci[2], Torsten Eymann[1]

[1] University of Bayreuth
[2] LABSS, ISTC-CNR Rome, Italy

Abstract. The Internet of Services is a term used to describe open global computing infrastructures in which an increasing number of services is made available to users through the Internet. Due to the openness, from the users point, the quality of the services offered can vary a lot and users have to concentrate on choosing the right ones. The choice of a good service thereby depends on the users' direct experience (Image), and their ability to acquire information (Reputation), which can be used to update their own evaluations. In this work, we present a set of simulation runs to explore the effect of reputation regarding services delivery in a Service network where information is asymmetrically distributed.

1 Introduction and Related Work

Businesses have to face several different challenges, when it comes to using Information Technology (IT). The increasing dynamism of markets leads to a continuous need for IT-Business-Alignment and the control of IT investments and resources. Thus, within the last years for every-day business the average operational expenses for IT (including the usage and maintenance of IT-infrastructures) have risen drastically, although actual prices for the single components decreased. The reason for this contrarian effect can be seen in the increased need for IT-resources like cpu power or storage capacity in order to cover the more and more computationally intensive IT tasks necessary for implementing new flexible business models within a short time. The main problem thereby is that the resources required need to be dimensioned to cover peak demand times, while only been sparsely used otherwise. Hence businesses face the problem that the IT-infrastructure costs for covering the peak times are disproportionate with regard to the average degree of capacity utilization of the resources.

The *Internet of Services* (IoS) describes a new computational paradigm, which envisions that in contrast to traditional models of web hosting where the web site owner purchases or leases a single server or space on a shared server and is charged a fixed fee, it leases the resources from an external provider and thereby substitutes the fixed costs by variable costs. The idea behind this vision is that if a company has to pay only for what it is using it can adapt its cost structure and will be able to economize, i.e. save money, while the providers offering

resources can benefit from economies of scale by using the same infrastructure to serve multiple clients [1]. The corresponding business models are twofold and can be summarized with the two terms *Software as a Service (SaaS)* and *Infrastructure as a Service (IaaS)*. SaaS describes the paradigm that providers offer their software products in an Internet environment that can be accessed at any time from any computer by the buyer of the service. The service sold thereby is an end-user-application that is restricted to what the application is and what it can do. Hence, the buyers neither know or control details of the underlying technology but only use the service as such. IaaS on the other hand is the hardware counterpart to SaaS. Thus in the IaaS business model case customers do not pay for services, but pay to use a shared infrastructure. In our view, SaaS and Iaas can build on top of each other, resulting in a scenario in which providers can play more than one role. This scenario will be used for our simulations. It will be explained in more detail in section 3.

In order to get such an environment running an efficient allocation mechanism is needed to match the demand and the supply of IoS resources - a market [2]. Markets have the advantage of collecting existing resource and service supplies and the corresponding demands and thereby usually achieve an even utilization by leveling heterogeneous user behavior. Like other utilities the services to be traded on those markets in huge numbers are of a simple nature. They are distinguishable by service quality characteristics, but convertible otherwise. This means that given equal characteristics competition will take place by signaling lower prices. However, keeping this ubiquitous vision of the IoS in mind, several questions occur, such as the question about the risks involved in IoS market transactions. Thus, it has to be ascertained that the bilateral economic exchange envisioned in IoS markets is very likely to involve risks, such as risks resulting from strategic- and parametric uncertainties [3].

Whereas the latter ones refer to environmental uncertainties that cannot (or only with a disproportionate effort) be reduced by the market participants, the strategic uncertainties concern the question of whether the transaction partners are willing to comply with what has been agreed on or not; and whether, if a transaction has had an adverse outcome, this was due to bad luck or bad intentions [4]. Thus, if a buyer does not receive the promised services or resources from the seller, it is often hard to judge whether the seller did not deliver intentionally, or whether the transaction failed, because the network broke down for example. This problem of strategic uncertainties is especially relevant for IoS markets, as in contrast to small communities where information about the transaction partners' trustworthiness can be conveyed by earlier experience with the potential transaction partners [5] or by other behavior-detection mechanisms [6], because IoS markets - as described above - have a large-scale nature which makes the mentioned mechanisms hardly applicable. In this paper we will focus only on strategic behavior.

One mechanism that nevertheless seems promising in this context and that we envision for reducing uncertainties and increasing trust in IoS markets is reputation, i.e. a judgment of transaction partners based on their past behavior.

The idea of dealing with reputation in service economies is not completely new. Alunkal et al. [7] use reputation information to select grid services, whereas Anandasivam and Neumann [8] use reputation information to fix prices for these services. Jurca and Faltings [9] use the quality of services to determine the price that has to be paid for a service. The quality of a service in turn is defined by reputation information. Finally, Silaghi et al. [10] review some additional approaches, but most from P2P field. In this paper we will analyze the effects of reputation in an IoS based on market coordination with the help of simulation.

In detail, we will start by deriving hypotheses concerning the influence of reputation on IoS transactions with the help of literature 2. Afterwards, we will introduce an IoS market scenario (see section 3) that will be used for the simulation to test the hypotheses. The simulation will then be conducted based on the market scenario and the hypotheses. The results of the comparative simulation runs are finally analyzed in section 4. The paper closes with a short summary of the findings and a conclusion (section 5).

2 Experimental Hypotheses

After this short introduction to the problem of reducing uncertainty and increasing trust in the IoS, in this section will will derive our hypotheses concerning the effects of reputation on the IoS with the help of literature. Thereby the main focus will be on social science and cognitive literature as within this area of research reputation and its effects have been discussed at length.

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

Image is a global or averaged evaluation of a given target on the part of an individual. It consists of a set of evaluative beliefs [11] about the characteristics of a target. These evaluative beliefs concern the ability or possibility for the target to fulfill one or more of the evaluator's goals, e.g. to behave responsibly in an economic transaction. An image, basically, tells whether the target is "good" or "bad", or "not so bad" etc. with respect to a norm, a standard, a skill etc.

In contrast *reputation* is the process and the effect of transmission of a target image. The evaluation circulating as social reputation may concern a subset of the target's characteristics, e.g. its willingness to comply with socially accepted norms and customs. More precisely, we define reputation to consist of three distinct but interrelated objects: (1) a cognitive representation, or more precisely a believed evaluation (any number of agent in the group may have this belief as their own); (2) a population-level dynamic, i.e., a propagating believed evaluation; and (3) an objective emergent property at the agent level, i.e., what the agent is believed to be as a result of the circulation of the evaluation [12].

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

But how do image and reputation effect uncertainty and trust in the IoS? In order to arrive at hypotheses to answer this questions we will briefly look at the general effects of reputation on transactions discussed in literature and abstract from these in a second step.

As argued by Conte and Paolucci [12] or Dellarocas [13] for example, one of the key functions of reputation is its impact for overcoming the challenges of moral hazard and adverse selection.

Moral hazard can be present any time two parties enter an agreement with one another. Each party in a contract may have the opportunity to gain from acting contrary to the principles laid out by the agreement. For example on eBay, the buyer typically sends the money to the seller before receiving the goods. The seller then is tempted to keep the money and not ship the goods, or to ship goods that are inferior to those advertised. The buyer thus has to take the risk of being cheated, because if he does not, no deal will take place. The two parties are trapped in a so called one-shot prisoner's dilemma (PD) with a one-sided exploitation in the case of advanced deliveries or payments. As shown by Ockenfels [14] for example, without any interference, in such a setting the participant reacting as second is always better of if he defects as his payoffs for the single transaction will be higher than in the cooperation case and he does not have to fear any future financial penalties as no record of his misbehaving is being kept. A possible solution to this problem is reputation, as with the help of reputation mechanisms the independent transactions can be linked as shown in figure 1.

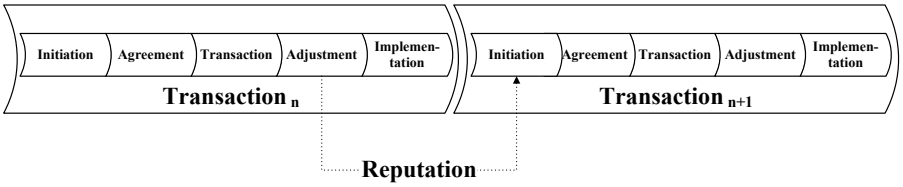


Fig. 1. Reputation as a linking mechanism between transactions

Reputation is the output of the adjustment phase of the first transaction and the input for the initiation phase of the next transaction. This "reputation history" is consequently a source of information that possible trading partners can use when estimating the trustworthiness and reliability of their counterparts and detecting defecting market participants. Consequently, reputation mechanisms can deter moral hazards by acting as sanctioning devices. If the community follows a norm that punishes traders with a history of bad behavior (i.e. by refusing to buy from them) and if the value punishment exceeds the gains from cheating, then the traders are all better off if they cooperate. Adverse selection is the second mechanism that can compromise cooperation in transactions. It is present in situations where sellers have information (about some aspects of their innate

ability, product quality, etc.) that buyers do not have (or vice versa), and i.e. information asymmetries exist. For example, a service provider has more information regarding the provided service than a potential service consumer. Thus, the consumer has to overcome this uncertainty before the agreement is signed. As shown by Akerlof this might result in a "market of lemons" [15]. Reputation mechanisms alleviate adverse selection issues by acting as a signaling device by soliciting and publishing the experience of consumers who have bought services from a certain seller [16].

Already these two short examples show that reputation has got several functions. First of all it works as an signaling device to distinguish between trustworthy and untrustworthy transaction partners. Furthermore, it changes the long term utility functions of the markets participant (by introducing potential losses in profit if being identified as cheater) and thereby encourages the transaction partners to cooperate. Due to the closeness of the IoS to the scenarios used in the papers mentioned above, we expect similar effects in IoS markets (like Eymann et al. [17]) and arrive at the following first hypotheses:

Hypotheses 1: Reputation reduces the uncertainty in the IoS by conveying cooperation.

For reasons of usability we have split this hypotheses in two sub-hypotheses that can be seen in the below. Thus we presume that first of all reputation will decrease the number of frauds as the shared image of the buyers helps them to identify cheater more easily (Hypotheses 1.1) and furthermore we surmise that the longer reputation information is used, the more information will be available so that the probability for defrauding actions to take place decreases over time (hypotheses 1.2):

Hypotheses 1.1: In the long run the fulfillment rate on SaaS and IaaS markets will be higher with than without reputation.

Hypotheses 1.2: If reputation is being used on SaaS and IaaS markets, in the long run the total fulfillment rate will exhibit an increasing (or at least constant) progression, but it will not decrease (except for short term variabilities).

Besides the first main hypotheses that reputation increases the number of transactions being fulfilled compared to a situation without reputation, in this paper we would like to investigate a second main hypothesis that is very closely related to the definition of reputation explained above. Thus, one of the main aspects of reputation is the circulation of information in a system. As it can be assumed that the further information spreads in a system the faster the system is penetrated with reputation information and the more information a single market participant can obtain about a potential transaction partner. Hence, if only image information is being used in a market for example and consequently the reputation spreading is limited very strongly all single buyers will have to

trade at least once with a certain seller before they can assess its trustworthiness react accordingly. In contrast if buyers share their images and consequently reputation information circulates through the whole system, buyers can obtain information about a seller faster and consequently the effects of reputation occur earlier.

Hypotheses 2: The further information can spread in a system, the faster reputation can take effect.

3 An Internet of Services Market Scenario

After having laid out and explained our hypotheses in the last section, in this section the market structure shall be looked at more closely.

3.1 The underlying Infrastructure

This section describes the model used in this simulation. The network representing an IoS is defined by a connected non-oriented graph, represented by a set of network sites $S = \{1, .., n\}$ and a set of links between the sites $L = \{\langle i_1, j_1\rangle, .., \langle i_m, j_m\rangle\}$. In addition, a failure probability f_{S_i} is defined for each node. When a failure occurs during simulation, the node is not able to answer any request or routing further messages. Which site will fail in each time tick, is chosen by chance due to the failure probability. Furthermore, on each node a set of the three different agent types, Complex Service Agents (CSA), Basic Service Agents (BSA) and Resource Agents (RA) is initialized. For more details on the agent functionality please see subsection 3.2. For each site the number of economic agents is $|CSA_{S_i}| \geq 0, |BSA_{S_i}| \geq 0, |RA_{S_i}| \geq 0$. A node without any associated economic agent is a router. Each link $\langle i, j\rangle$ between two nodes has a certain bandwidth. A higher bandwidth leads to an increased data transfer. In our simulation model, the bandwidth is biased, which means that a link is defined or not. Nodes, which are not linked directly can be addressed through a routing table that is calculated by a common routing algorithm, the Dijkstra algorithm [18].

The entry point to the simulation is the CSA. The CSA has to fulfill an external generated demand. In our simulation the demand is generated with an uniformly distributed interval between to demand arrivals. The kind of basic service the CSA has to buy is also given by demand generation. The BSA on the other hand has to compose different resources by a certain combination. The demanded resource bundle can differ between the BSA-types.

3.2 The Market Structure

The structure of the market that is used for the IoS simulation experiments was derived from the ideas of the EU-funded project CATNETS[3]. The overall structure consists of four types of players, which act on two interrelated markets, namely the SaaS and IaaS that can be seen in figure 2. Hence, our approach does not only incorporate services as basic units provided to consumers within the IoS system, but also defines a IaaS market trading the actual computational resources needed for implementing those services.

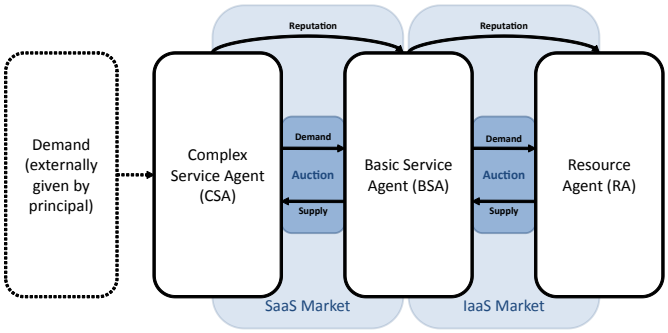


Fig. 2. The Market Structure

Starting from the right side of the figure these markets and the corresponding participants acting on the markets are: (1) the IaaS market - which involves trading of computational and data resources, such as processors, memory, etc. between RA (sellers of the resources) and BSA (buyers of the resources), and (2) a SaaS market – which involves trading of basic application services between BSA as sellers and the CSA as buyers of basic services[4].

In detail the scenario for trading on the markets works as follows: If a CSA is given a demand for a complex service externally by its principal it will try to satisfy this demand by breaking the complex service down into several basic services and buying the basic services required on the SaaS market from the BSA. In order to be able to buy basic services, the CSA is being given an initial budget that it can spend. After a deal on the SaaS market is closed, the CSA will

[3] For more information on the CATNETS project see http://www.catnets.uni-bayreuth.de/

[4] The distinction between resource and service is necessary to allow different instances of the same service to be hosted on different resources. It also enables a given service to be priced based on the particular resource capabilities that are being made available by some hosting environment.

pay the BSA the money for the services. In order to be able to "produce" basic services, the BSA needs resources which it can buy on the IaaS market from the RA. It thereby can invest the money it got from the CSA for the transaction on the SaaS market (optional: and other budget it may have saved from earlier successful transactions). Again, if a deal is closed (on the IaaS market) the BSA pays the RA in advance. The RA is the last in the chain and can finally decide, whether it delivers the sold resources to the BSA or not (i.e. it can decide whether it defects or not). In case it defects, the BSA will not get the resources and consequently will not be able to "produce" the basic services that it has sold the CSA and will fail its delivery. In case the RA delivers the promised resources, the BSA can and will deliver the basic services to the CSA (i.e. it will never defect). If the RA fulfills the invocation, it is locked for a certain time. As feedback to the transactions, one-sided reputation is given. Thus, the CSA will rate the BSA and the BSA the RA. As the two markets are interrelated allocating resources and services on one market inevitably influences the outcome on the other market. Consequently, the BSA risks a negative reputation in case it cannot deliver its services to the CSA in time. Therefore, it has incentives to choose a reliable RA as a trading partner and consequently it is highly interested in the reputation information on the RAs.

After this brief explanation of the market structure, now the negotiation process to be used for all SaaS market transactions as well as for all IaaS market transactions shall be explained in more detail.

3.3 The Negotiation Protocol

Within the simulations we will concentrate on the English Auction protocol as negotiation mechanism, both on the IaaS and the SaaS market. The assumed market structures and the fact that services and resources are sold and not requested determines, in our point of view, the English Auction protocol. Each selling agent (that are BSA on the SaaS market and RA on the IaaS market) sells its own service, that means the agent fulfills also the role of an auctioneer. As using a time discreet simulation environment (see section 4) at each time tick exactly one agent is able to decide whether to start an auction or not. As a consequence, the buyers need to decide, whether they buy from the agents offering services/resources at a specific point of time, or whether they wait some more time until the next (potentially reputation-wise better) seller offers its product. However, the more time passes by, the risk of not getting any of the needed services/resources increases.

The seller (and auctioneer) proposes an auction and all agents, which are interested register for participating. The call for bids messages are sent to all participants, which can then place their bids. The increasing price results in an outdropping of bidders of the auction. In the end, the last remaining buyer wins the auction and has to pay the second-highest bid.

After the auction, the seller sends a message with the winning bid and the winner to all participating agents. They can use the information for comparison with their own bidding, and thus learn towards a better strategy for the next

auctions. In this simulation the learning strategy is simplified, such that agents are just increasing their reservation price when loosing an auction and decreasing it otherwise.

3.4 The Reputation Mechanism

As we want to analyzed the implications of reputation on distributed IoS markets, besides the market described above, a reputation mechanisms applicable to distributed infrastructures is needed. Even if a centralized model is easier to use for the simulation designer and the reputation information is spreading faster, a decentralized model, like the one of Alfarez Abdhul Rahmen and Stephan Hailes [19–21] is more realistic in this context. Their model that pursues a perspective different from M. Schillo's "Trust-Net"[22, 23], is directly related to internet-based MAS and is supposed to help to implement trust as the basis of informal-, short-term-, or commercial ad-hoc transactions. Therefore they propose that every agent carries along a network of trust relationships in a database, hence information are stored decentralized. Abduhl-Rahman and Hailes define a "trust-relationship" as a vectored connection between exactly two entities, which in some circumstances can be transitive. In this way they distinguish between direct trust relationships ('Alice trusts Bob.') and recommender trust relationships ('Alice-trust-Bob recommendations about the trustworthiness of other agents'). An interesting contrast to other formalizations lies in the fact that due to the qualitative nature of trust, Abduhl-Rahman and Hailes do not work with probability values or the $[-1; 1]$ interval, but interpret trust and distrust as a condition (and not as a continuum) and therefore propose a use a multi-context implementation, in form of discrete values that are related to certain trust categories ("Alice trusts Bob, concerning 'table'-transactions. However, she does not trust him when it comes to 'chair'-transactions."). The discrete values used can be seen in table 1.

As a result, Abduhl-Rahman and Hailes define reputation as a "troika" $(agent-ID, Trust-Category, Trust-Value)$. Each agent stores such reputation information in its own data-base and uses it to articulate recommendations.

The core of Abduhl-Rahman and Hailes' papers allegorize their thoughts about a recommendation protocol that can be used to communicate recommendation requests and statements as well as updating inquiries within the MAS. In the protocol, a recommendation request, for example, is passed on until one or more agents are found which can give information for the requested category and which is trusted by the penultimate agent in the chain. Based on this idea Abduhl-Rahman and Hailes propose a mathematical algorithm for the rating phase in which the requesting agent can use the following equation to calculate the trustworthiness of a recommendation. For $tv(R_x)$ as the recommender trust value of the different recommendations of the involved agents and $rtv(T)$ as the trust value articulated by the last agent[5] the trustworthiness result from the following equation:

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

Value	**Significance for direct trust relationship**	**Significance for recommender trust relationship**
−1	Distrust - completely untrustworthy	Distrust - completely untrustworthy
0	Ignorance - cannot make trust-related judgment about entity	Ignorance - cannot make trust-related judgment about entity
1	Minimal - lowest possible trust	"closeness" of the recommender's judgment to own judgment about trustworthiness
2	Average - mean trustworthiness (most entities have this trust level)	
3	Good - more trustworthy than most entities	
4	Complete - completely trust this entity	

Table 1. Discrete trust value after [20, p. 53]

$$tv_r(T) = \frac{tv(R_1)}{4} * \frac{tv(R_2)}{4} * ... * \frac{tv(R_n)}{4} * rtv(T) \quad (1)$$

This qualitative and at the same time algorithmic approach made the model of Abduhl-Rahman and Hailes interesting to the work presented in this paper and therefore we decided to implement it for our simulations.

For the actual implementation of the just described reputation mechanism we focused on a paper by Pinyol et al. [24]. In their implementation proposal Pinyol et al. strictly concentrate on the mechanism and only extend it slightly in one point. Hence, in order to ease the adaption of the agent strategies they transfer the discrete values derived from other agents into probabilistic sets between 0 and 1. This is done by setting the value for completely trustworthy behavior to 1 and the one for completely untrustworthy behavior to 0 and by then arranging the remaining values accordingly in order to be able to calculate the deficiency probability.

3.5 Using reputation information during the negotiation process

The buying agent can use reputation information when a call for bids arrives. It has to decide based on its own image if it participates in the auction. Is the own information about the seller not strong enough or is there even no previous experience with the target, the agent can ask neighbored agents for their estimation. The size of the neighborhood can be parametrized through the limitation of hops agents are able to send their call. This issue is also discussed in section 4.

In case, the agents opinion is positive (optionally after retrieving shared image information from other agents), the agent will participate in the proposed auction. For that, the image, based on the continuous spectrum $image = [0..1]$, has to exceed a threshold defined by the agent strategy.

To keep a trust and reputation system alive, the agents have to adapt their own opinion about a target after having own experience or when receiving third party's information. After a transaction has been finished (from CSA's or BSA's) point of view, the agent adapt their own image about the transaction partner in dependence of the transaction's outcome. Furthermore, the image threshold is increased or decreased, which means for following transactions that the agent accepts more or less risk. Additionally we assume that agents share their image, as they are beneficiaries of the trust and reputation system.

4 Analysis of Simulation Result

To test our hypotheses, we use a simulation environment based on the Multi-Agent Simulation Toolkit Repast[6]. The underlying network consists of 100 nodes, connected in an Internet-like way without any clusters or heavy-tailed elements. Further, 160 CSA, 200 BSA and 40 RA have been initialized in the system. At the initializing phase the RA are parametrized concerning whether they will cheat (that means not to answer their invoking requests) or not. This decision is a binary one, hence, either the agent is always cheating or it is never cheating. The rate of cheating agent is for all following simulation runs constant at 20 percent (i.e. eight RA). The cheating agents are distributed like all other agents arbitrary on the nodes.

The main task of the reputation mechanism will be to identify the cheating agents. To test our hypotheses we will limit the decentralized reputation mechanism in its spreading range. The availability of the nodes is also limited. For an agent starting an auction this limitation means that it can not propose its auction to all agents in the system. Only its neighbors are addressed. This assumption seems to be quite reasonable, because sending messages through the network is time- and money-consuming. So sending messages just to an sufficient number of agents is reasonable for sellers, due to the less time auctions remain. Buyer agents, on the other hand, have an interest in receiving as many auction proposals as possible, as they might overcome missing information.

To test our hypotheses we will use the *characteristic* fulfillment rate. During simulation (in each tick) every agent logs its accumulated successful transactions and its failed transactions. The characteristic denotes the mean value over all agents, which had at least one interaction. A transaction is counted as successful, if and only if a fulfillment message arrives within the timeout at the buyer agent. If the timeout expires the agent will record a failed interaction. Note: For mathematical reasons the ratio is more volatile at the beginning of the simulation. After some interactions a single failed or successful interaction does not effect

[6] Repast Organization for Architecture and Design, Available at http://repast.sourceforge.net/ (2008)

the characteristic so strong compared with the beginning of the simulation. Because of this artefact the first few hundred time ticks should not be considered within the interpretation.

To test Hypotheses 1 that reputation reduces the uncertainty in the IoS by conveying cooperation we compare two typical simulation runs. During the first simulation run the trust and reputation model from Abduhl-Rahman and Hailes (see section 3.4) is used to spread agents' images. In figure 3 the blue lines represent the fulfillment rate over time with shared image usage. The agents are able to ask for opinions for a given target over two hops. That means all agents on the same node and all agents on nodes within the hop limit are asked for their opinion. Thereby, as only agents of the same type can have own experience with the target agent, only this subset of agent will answer the request.

The red graphs on the other hand represents the fulfillment rate without trust and reputation usage. With some variation at the beginning this value oscillates around 80 percent. This value is exactly the one we expected, because, like mentioned above, 20 percent of RA are cheating.

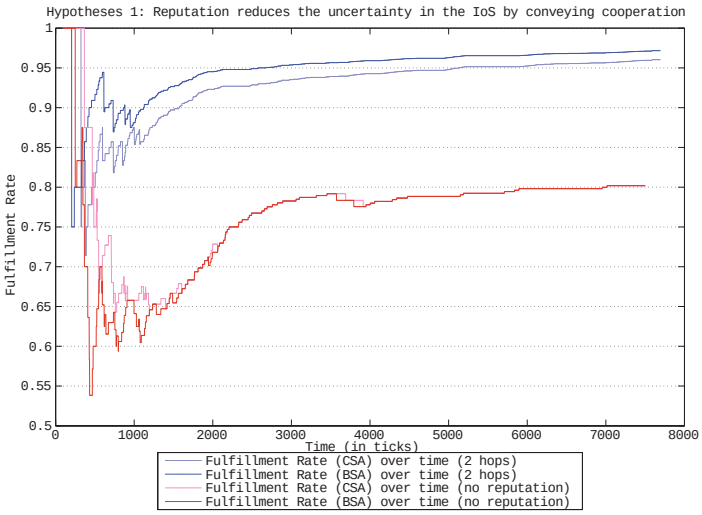


Fig. 3. Simulation with shared image usage compared to no usage of trust and reputation models

Comparing the two simulation outcomes, we can see that in the second case (when using no trust and reputation model) even later than in the other case cheating agents are involved. In the first case after about 1200 ticks no cheating agent is involved in a transaction, that means all cheating resource agents are identified. If they propose an auction no BSA bids on that auction. The difference between the corresponding BSA and CSA lines can be explained by the different

timeouts. Sometimes the timeout is as long, such that BSA if they by a service from cheating RAs, have enough time to buy a second resource to fulfill their own promise on the SaaS market. That is why the rates are not equal over time.

Regarding Hypotheses 1.2 we can see that the fulfillment rate is still increasing over simulation time. But even in longer simulation runs it will never reach the value 1, due to the failed transaction at the beginning of the simulation when the trust and reputation system has yet not been filled with enough evaluations.

To test hypotheses 2, we will compare the case using two hops to announce and to ask for own experiences with the case of sending no reputation messages to other hops. The latter case means that only agents on the same node can be asked for their experiences with the target node. The outcome of the simulation can be seen in figure 4.

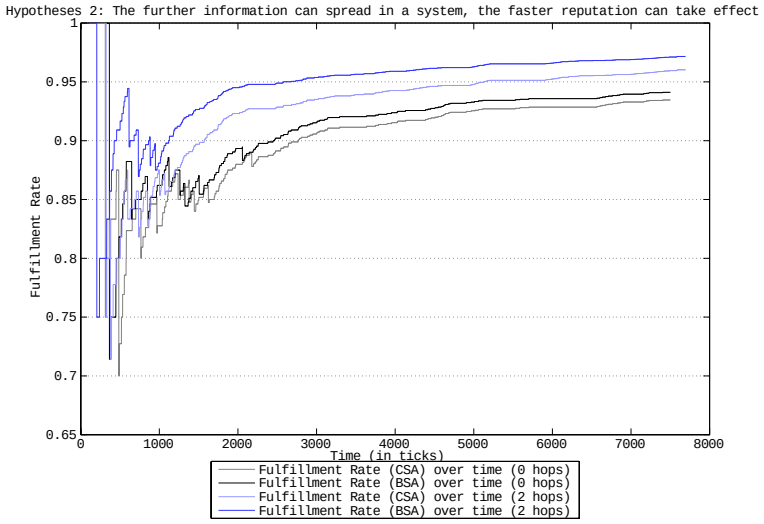


Fig. 4. Comparison between experiments with reputation spreading to other nodes and agents not sending reputation messages to other nodes

Also in this simulation experiment we can notice, beside the oscillation at the beginning of simulation run (see explanation above), an increased fulfillment rate when spreading the trust and reputation information wider in the system. Each agent is able to update its own image with more information from other nodes. The information spreading is faster what can be seen in the fact that in the first case (blue lines) from tick 1200 on no cheating agent is involved any more, but in the latter case (black lines) until tick 2000 some interactions fail. In addition, the fulfillment rate increases faster the further information is spread. This can be seen in the steeper elevation of the two-hops-curve.

5 Conclusion and Future Work

We can conclude that we introduced two main hypotheses in this paper. The first one is that reputation might reduce uncertainty in an IoS. In our simulation on a typical Internet structure and different service types running on these nodes support this hypotheses. More jobs can be fulfilled correctly, if a trust and reputation mechanism is introduced. The second hypotheses regards the information spreading within the reputation system. Enforced to use a decentralized mechanism we have to find a suitable tradeoff between time-consuming soliciting reputation information and the better information of each agent. A centralized mechanism is not very common to use in an Internet-like network structure due to the loss of relevance in reality. Consequently such as the first hypotheses, the second one can be substantiated as well.

For future work we will have to consider different aspects of our results in detail. The connection between the radius of information spreading and the radius of auction information for example. Further we should drop the assumption of using the English Auction protocol as the one negotiation protocol. With regard to [25] we should be able to test some of the hypotheses the authors propose there.

Further hypotheses to test might be the fairness between the agent types. Do honest agents perform better regarding their utility function than cheating agents? Regarding the environment we should be able to provide a real world prototype which can be improved with reputation mechanism. Last but not least, the interrelation of the two markets that exhibit a supply-chain-like relationship needs further attention.

Acknowledgment This work results from the EU-funded project eRep – Social Knowledge for e-Governence. For more information please visit the project web site on http://megatron.iiia.csic.es/eRep/

References

1. Carr, N.G.: Does IT matter? Harvard Business School Press, Boston, MA (2003)
2. Streitberger, W., Hudert, S., Eymann, T., Schnizler, B., Zini, F., Catalano, M.: On the simulation of grid market coordination approaches. Journal of Grid Computing; Special Issue on Grid Economics **6**(3) (2008) 349–366
3. Voigt, S.: Institutionenökonomie. Neue Ökonomische Bibliothek. UBT Verlag (2002)
4. Güth, W., Kliemt, H.: Evolutionarily stable co-operative commitments. Theory and Decision **49** (November 2000) 197–221
5. Axelrod, R.: The Evolution of Cooperation. Basic Books, New York (1984)
6. Frank, R.H.: If homo economicus could choose his own utility function, would he want one with a conscience? American Economic Review **77**(4) (September 1987) 593–604
7. Alunkal, B., Veljkovic, I., von Laszewski, G., Amin, K.: Reputation-based Grid Resource Selection. In: Workshop on Adaptive Grid Middleware, New Orleans, Louisiana, AGridM 2003 (September 2003)

8. Anandasivam, A., Neumann, D.: Reputation-based pricing for grid computing in e-science. In: 16th European Conference on Information Systems (ECIS 2008). (2008)
9. Jurca, R., Faltings, B.: Reputation-based service level agreements for web services. In: Service Oriented Computing (ICSOC - 2005). Volume 3826 of Lecture Notes in Computer Science. (2005) 396 – 409
10. Silaghi, G.C., Arenas, A.E., Silva, L.M.: Reputation-based trust management systems and their applicability to grids. Technical Report TR-0064, Institute on Knowledge and Data Management Institute on System Architecture, CoreGRID - Network of Excellence (February 2007)
11. Miceli, M., Castelfranchi, C.: The role of evaluation in cognition and social interaction. In Dautenhahn, K., ed.: Human cognition and social agent technology. Benjamins, Amsterdam (2000)
12. Conte, R., Paolucci, M.: Reputation in Artificial Societies: Social Beliefs for Social Order. Springer (October 2002)
13. Dellarocas, C.: Reputation Mechanisms. Handbook on Economics and Information Systems. Elsevier Publishing (2006)
14. Ockenfels, A.: Reputationsmechanismen auf internet-marktplattformen: Theorie und empirie. Zeitschrift für Betriebswirtschaft **73**(3) (2003) 295–315
15. Akerlof, G.A.: The market for "lemons": quality uncertainty and the market mechanism. The Quarterly Journal of Economics **84**(3) (August 1970) 488500
16. Jurca, R.: Truthful Reputation Mechanisms for Online Systems. PhD thesis, University of Lausanne (2007)
17. Eymann, T., König, S., Matros, R.: A framework for trust and reputation in grid environments. Journal of Grid Computing; Special Issue on Grid Economics **6**(3) (2008) 225–237
18. Dijkstra, E.: A note on two problems in connexion with graphs. Numerische Mathematik **1**(1) (December 1959) 269–271
19. Abdul-Rahman, A., Hailes, S.: A distributed trust model. In: NSPW '97: Proceedings of the 1997 workshop on New security paradigms, New York, NY, USA, ACM (1997) 48–60
20. Abdul-Rahman, A., Hailes, S.: Using recommendations for managing trust in distributed systems. In: IEEE Malaysia International Conference on Communication. (1997)
21. Abdul-Rahman, A., Hailes, S.: Supporting trust in virtual communities. In: HICSS. (2000)
22. Schillo, M., Funk, P., Rovatsos, M.: Who can you trust: Dealing with deception. In: Proceedings of the Second Workshop on Deception, Fraud and Trust in Agent Societies, Seattle, USA. (1999) 95–106
23. Schillo, M., Funk, P., Rovatsos, M.: Using trust for detecting deceitful agents in artificial societies. Applied Artificial Intelligence **14**(8) (2000) 825–848
24. Pinyol, I., Sabater-Mir, J., Cun, G.: How to talk about reputation using a common ontology: From definition to implementation. (2007) 90—101
25. König, S., Hudert, S., Eymann, T., Paolucci, M.: Towards reputation enhanced electronic negotiations for service oriented computing. In: Proceedings of the CEC/EEE 2008, Washington, DC. (2008)

Reputation and Uncertainty. A fairly optimistic society when cheating is total

Walter Quattrociocchi and Mario Paolucci

LABSS., ISTC-CNR Rome, Italy

Abstract. In an uncertain world, humans or artificial agents, to cope with uncertainty, need to communicate and share information to increase the number of their experiences, and consequently their possibility of success. The information shared by agents in a society can have different nature: accepted evaluations (Image) or reported voices (Reputation). In this work we model a simulative context where information acquisition is strongly affected by false information; in the experiments presented, performed on the RepAge Platform which is a computational module for the management of reputational information, agents can lie or report others' lies. In this work we explore the effect of informational cheating under extreme setting, focusing to study from one hand, the effect of cheating on the quality achievement when the society is composed by a large amount of liars, and on the other hand, the beliefs formation and revision dynamics when the informational domain is not reliable. Information accuracy has effect on the market in relation with its trustworthiness; if social information is not reliable communication loses its importance.

1 Introduction

In this paper, evaluation dynamics which are essential for intelligent agents to act adaptively in a virtual society will be examined through multi agent based simulations. The decision making process in a socially situated agent requires specific cognitive capacities (individual and social intelligence) to cope with uncertainty. A model for describing dependencies of different informational domains among agents endowed with different behaviors (honest and liars) was applied to a computer simulation study for partner selection dynamics in a basic market. Information exchange is a combined social function in which individuals request, provide, and exchange information with the goal of reducing uncertainty and plays a basic function for formation of opinions.

2 Repage and the Underlying Theory of Reputation

Repage is built upon a theory of reputation originally presented in [2]. Its main characteristic is the fundamental distinction in social evaluations used in decision making process, i.e that between *image* and *reputation*. Both image and reputation are social evaluations concerning other agents' attitudes toward socially

desirable behavior; they exist on both an individual and a social level, and may be shared by a multitude of agents. We call image the evaluation that the agent actually believes; we call reputation the information the agent considers to be spread as a reported voice. Consequently, reputation differs from image because is true when it is actually spread, not when it is accurate; it does not bind the speaker to commit himself to the truth value of the evaluation conveyed, but only to the existence of rumours about it. Unlike ordinary communication and deception, reputation implies neither personal commitment nor responsibility. Hence, to acknowledge the existence of a reputation does not imply to accept the evaluation itself. For instance, agent A might have a very good image of agent B as a seller, and at the same time accept that it is said that agent B is a bad seller.

Reputation co-evolved with human language and social organization as a multi-purpose social and cognitive artifact [3]. It provides incentives for cooperation and norm abiding, while discouraging defection and free-riding by allowing retaliation against transgressors and enforcing cooperation at the level of information exchange. In this paper we will discuss the role of information trustworthiness in partner selection within a multi agent system simulating a market place. The various aspects of this process are explored with the Repage framework to simulate the effect of image and reputation in a marketplace. Repage [10] is a computational system based on the above-mentioned theory of reputation. For a detailed explanation of Repage's mechanisms, we refer to [10] and [5]. The Repage module is implemented in Java and is freely available as a sourceforge project.

Repage provides evaluations on potential partners and is fed with information from others and outcomes from direct experience. To select good partners, agents need to form and update their own social evaluations; hence, they must exchange evaluations with one another. If agents transmit only believed image, the circulation of social knowledge would be bound to stop soon. But in order to preserve their autonomy, agents need to decide whether to share or not other 's evaluations of a given target. If agents transmit others' evaluations as if these evaluations were their own, without the possibility of choosing whether merge both types of evaluations or not, they would be no more autonomous. Hence, they must

- form both evaluations (image) and meta-evaluations (reputation), keeping distinct the representation of own and others' evaluations, before
- deciding whether or not to integrate reputation with their own image of a target.

Unlike current systems, Repage allows agents to transmit their own image of a given target, which they hold to be true, or to report on what they have heard about the target, i.e. its reputation, whether they believe this to be true or not. Of course, in the latter case, they will neither commit to this information's truth value nor feel responsible for its consequences. Consequently, agents are expected to transmit uncertain information, and a given positive or negative

reputation may circulate over a population of agents even if its content is not actually shared by the majority.

3 Related Works

Reputation models abound in the field of agent based systems; we mention [12, 8] (for a comparison of these models from the point of view of information representation, refer to [9]). All of these models are essentially monodimensional, in the sense that image and reputation are made to collapse. As a consequence, there is no way to distinguish a high responsibility versus a low responsibility evaluation which instead is central to our theoretical analysis. Repage is then the only system able to provide insight on how these process may develop.

Simulations based on Repage have already been presented in [11] and compared with results obtained in a simpler system, implemented on NetLogo, and based on image only. In [5], the importance of retaliation in Repage is discussed in connection with evaluations implying high and low responsibility (i.e image and reputation). In [6] the authors show that a market with exchange of reputation exhibits characteristics that are absent in a market with exchange of image only; unlike the fluctuation found with image only, a stabilizing effect of reputation is visible in the discovery of a critical resource, information about good sellers. In this work we present a new set of simulations in extreme cheating conditions.

According to uncertainty reduction theory [1] which defines the goal of reducing uncertainty as a central motive of communication, we argue not only that a large amount of information allows uncertainty reduction, but also that different sources of information contribute to uncertainty reduction and consequently to cultural formation.

4 Experiments

4.1 Design of the Experiment

The market has been designed with the purpose of providing the simplest possible setting where information is both valuable and scarce. It includes only two kind of agents, the buyers and the sellers. All agents perform actions in discrete time units (turns from now on). In a turn, a buyer performs one communication request and one purchase operation. In addition, the buyer answers all the information requests that it receives. Goods are characterized by an utility factor that we interpret as quality (but, given the level of abstraction used, could as well represent other utility factors as quantity, discount, timeliness) with values between 1 and 100.

Sellers are characterized by a constant quality, drawn following a stationary probability distribution, and a fixed stock, that is decreased at every purchase; they are essentially reactive, their functional role in the simulation being limited to providing an abstract good of variable quality to the buyers. Sellers exit the

simulation when the stock is exhausted and are substituted by a new seller with similar characteristics but with a new identity (and as such, unknown to the buyers). This continuous seller update characterises our model, for example in comparison with recent work as [4], where both sellers and buyers are essentially fixed.

The disappearance of sellers makes information necessary; reliable communication allows for faster discover of the better sellers. This motivates the agents to participate in the information exchange. In a setting with permanent sellers (infinite stock), once all buyers have found a good seller, there is no reason to change and the experiment freezes. With finite stock, even after having found a good seller, buyers, should be prepared to start a new search when the good seller's stock ends.

At the same time, limited stock makes good sellers a scarce resource, and this constitutes a motivation for the agents not to distribute information. One of the interests of the model is in the balance between these two factors.

There are five parameters that describe an experiment: the number of buyers, the number of sellers, the stock for each seller, the distribution of quality among sellers, and the percentage of cheaters. We define the two main experimental situations as L1 where there is only exchange of image, and L2 where both image and reputation are used.

4.2 Agents Decisions and Actions

The agents decision making procedure is the point where reputation is put to work. From the seller side, this procedure is limited to sell the products that buyers require and to disappear when the stock gets exhausted. From the point of view of the buyers, at each turn they have to ask one question to another buyer and buy some item from a seller. They may also answer a question from other buyers. Buyers actions are:

- Buying Action
- Asking Action
- Answering Action

Buying action In this action the question is: which seller should I choose? The Repage system provides information about image and reputation of each one of the sellers. The easiest option would be to pick the seller with better image, or (in L2) better reputation if image is not available. We set a threshold for an evaluation to be considered good enough to be used to make a choice. In addition, we keep a limited chance to explore new sellers, controlled by a system parameter. Notice that image has always priority over reputation, since image imply an acknowledgement of the evaluation itself while reputation only an acknowledgement of what is said. For a more formal description of the decision procedures refer to [5].

Asking action As in the previous action, the first decision is the choice of the agent to be queried, and the decision making procedure is exactly the same as for choosing a seller, but dealing with images and reputation of the agents as informers (informer image) instead of as sellers. Because of the fear of retaliation, sending an image will take place only when an agent is very confident of that evaluation, in the sense of the RepAge strength parameter included in every evaluation.

Once decided who to ask, the kind of question must be chosen. We consider only two possible queries: Q1 - Ask information about a buyer as informer (basically, how honest is buyer X as informer?), and Q2 - Ask for some good or bad seller (for instance, who is a good seller, or who is a bad seller?). Notice that this second possible question does not refer to one specific individual, but to the whole body of information that the queried agent may have. This is in order to allow for managing large numbers of seller, when the probability to choose a target seller that the queried agent have some information about would be very low. If Q1 is chosen, buyer X as informer would be the less known one, that is, the one with less information to build up an image or reputation of it.

Answering action Let agent S be the agent asking the question, R the agent being queried. Agents can lie, either because they are cheaters or because they are retaliating. When a buyer is a cheater whatever information being answered is changed to its opposite value. Retaliation is accomplished by sending inaccurate information from the point of view of the sender (for instance, sending *Idontknow* when really it has information, or simply giving the opposite value) when R has a bad image of S as informer. In L1 retaliation is done by sending an *Idontknow* message even when R has information. This avoids possible retaliation from S since an *Idontknow* message do not imply any commitment. If reputation is allowed, (L2) retaliation is accomplished in the same way as if the agent were a liar, but converting all image to send into reputation, in order to avoid as well possible retaliation from S. We present as algorithm 1 an example of the decision procedure used to answer questions of type Q1 in the presence of reputation (L2).

Algorithm 1 Answering Q1 Decision Procedure R in L2

1: ImgX := Get image of agent X as informant;
2: if ImgX exists and strength(ImgX) $\geq$ thStrength then send ImgX to agent S, END;
3: else convert ImgX to RepX and send RepX to S, END
4: if ImgX does not exist then RepX := Get reputation of agent X as informant;
5: if RepX exists then send RepX to S, END;
6: if RepX does not exist then send *Idontknow* to agent S;

4.3 Simulative Scenarios

The decision making process in a socially situated agent requires specific cognitive capacities (individual and social intelligence) to deal with uncertainty. In this work, a model for describing dependencies of different informational domains among agents endowed with different behaviors (honest and liars) was applied to a computer simulation study for partner selection dynamics in a basic market. In the system, a double process of interaction should be described to allow for partner selection in a virtual market place. On one hand, opinion dynamics (social evaluation) are derived from lower-level phenomena (individual evaluation); on the other hand, partner selection is expected to be affected by the kind of information circulating. The experimental session presented in this paper investigates the role played by different kinds of information in beliefs formation on informational cheating. This work starts from an established property of reputation to reduce uncertainty (see [7]); the investigation addresses the differentiation of evaluations with respect to the information availability.

All the simulative sessions are centered on the investigation of the global system behavior when agents' informational domain contains almost false information or in the opposite case with only true information available. To this purpose, we examine simulations where the amount of informational cheaters is very low or very high, as shown in **Table 1**. For each experiment ten runs are performed.

Table 1. Experiments Settings

Experiment	Sellers	Buyers	Cheaters	Level	Stock	Good Sellers	Bad Sellers
$0CL1$	100	15	0%	L1	50	5%	10%
$0CL2$	100	15	0%	L2	50	5%	10%
$90CL1$	100	15	90%	L1	50	5%	10%
$90CL2$	100	15	90%	L2	50	5%	10%

4.4 Hypotheses

In this section we present a brief description of our experimental hypotheses:

- in social and economic exchange, partner selection is fundamental to increase chances of cooperation and quality of products exchanged.
- Image is characterized by two consequences: it is either (a) followed by retaliation, which partially neutralizes the good effect of partner selection, or (b) causing only tested information to be spread, which reduces the quantity of information available to agents.
- In the reputation condition, instead, partner selection is associated with less retaliation and more, although possibly uncertain, information circulate into the system.

– With reputation there is a larger quantity of information circulating in the system and we expect uncertain evaluations to decrease due to a larger quantity of information, and because meta-beliefs allow for a more synthetic description of a target.

To check these hypotheses, we designed two experimental conditions, with image only (L1) and with both Image and Reputation (L2). We explore several values of the parameters in order to study the respective impact of the two conditions. Given the hypotheses formulated above, we expect that:

– there is an initial advantage of L2 over L1, that is, L2 grows faster in average quality.
– L2 performs better as a whole, that is, the average quality at regime is higher than L1. Note that to obtain this result we are hardwiring a limitation in image communication, based on the theory that foresees large amounts of retaliation against mistaken image communications but not on the reputation side.
– L2 presents a smoother trend in the discovery of good informers than L1 via the exchange of meta-beliefs.
– L2 presents an higher uncertainty reduction than L1 for in an informational domain where both types of evaluation circulate.

5 Results

In this section simulation results are presented in an overview and then charts are shown and discussed in relation to the hypotheses.

In the following tables all the statistical values are computed on the last 20 turns for each experiment, when the system has reached a stable state. In the first table (**Table 2**) we report a summary of average values and standard deviations for each simulation scenario. We report average quality over agents in repeated experiments, resulting from the quality of individual contracts. GS stands for "good sellers discovered" and gives the number of good sellers found out in the current turn; U gives the number of answer based on uncertain evaluations, given in the current turn.

The second table (**Table 3**) presents full information and statistics on Quality, Good Sellers discovery and Uncertainty, reports the following standard statistical values: **MAX** as the maximum value, **MIN** the minimum value, **AVG** the average, **STD DEV** indicating the standard deviation value, **VAR** as the variance, **DEV ABS** the absolute deviation, **ASIMM** as asymmetry value, **KOURT** the kourtosis coefficient and **AUTO-COR** the auto-correlation coefficient value for each experiment. When there are only honest agents the performances are similar both in L1 and L2 settings.

Quality is slightly higher without cheaters, but there is no difference between L1 and L2 in both settings. Good Sellers discovery is higher in $0CL1$ and maintains the same values in all the other cases. The higher value of GS is a sign of an intense search activity caused by a lower amount of available information.

Uncertainty is reduced by reputation in both conditions, with many and few cheaters ($0CL2$ and $90CL2$). In particular looking at Quality values which reach the same levels without presenting any relevant difference. The Good Sellers Discovery trend presents some differences in the two scenario; we believe this is caused to the high number of good sellers and finite stock. Uncertainty decreases with reputation spreading in both extremes.

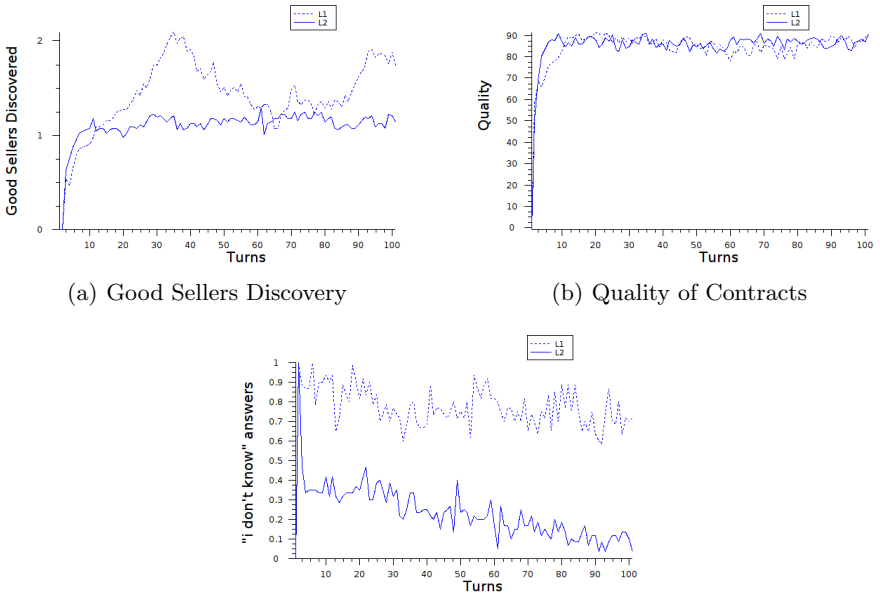
Table 2. Simulation Parameters Resume

	Q	GS	U
0CL1	86,93(1,47)	1,62(0.19)	0,71(0.06)
0CL2	86,41(1,69)	1,12(0,044)	0,09(0,02)
90CL1	83.72(2.2)	1.1(0.06)	0.7(0.06)
90CL2	82.53(2.35)	1.07(0.04)	0.19(0.03)

Table 3. Simulation Results Resume

VAL	MAX	MIN	AVG	STD DEV	VAR	DEV ABS	ASIMM	KOURT	AUTCOR
0CL1									
Q	89,56	83,56	86,93	1,88	3,53	1,47	-0,45	-1,01	0,16
GS	1,9	1,26	1,62	0,22	0.05	0.19	-0,27	-1,63	0,85
U	0,88	0,58	0,71	0,08	0.007	0,06	0,48	0,69	0,17
0CL2									
Q	90,11	82,71	86,41	2,11	4,46	1,69	-0,24	-1,01	0,47
GS	1,21	1,05	1,12	0,05	0,002	0,044	0,36	-1,39	0,33
U	0,16	0,033	0,09	0,03	0.0013	0,02	-0,12	-0,77	0,07
90CL1									
Q	87.76	75.86	83.72	2.84	8.06	2.20	-0.81	0.61	0.41
GS	1.21	0.96	1.10	0.07	0.005	0.06	-0.08	-1.17	0.79
U	0.85	0.6	0.70	0.07	0.005	0.06	0.14	-1.18	0.20
90CL2									
Q	86.08	75.58	82.53	3.001	9.009	2.35	-0.77	-0.47	-0.11
GS	1.18	0.96	1.07	0.05	0.002	0.04	-0.014	-0.62	0.53
U	0.28	0.1	0.19	0.04	0.002	0.03	0.066	-0.38	-0.19

Figure 1a shows the good sellers discovery in absence of informational cheaters for 100 simulation turns ($0CL1$ and $OCL2$); the curves for L1 and L2 present fairly different behavior. In L1 the peaks of each wave represent phases of high search, when all agents actively explore different sellers until they find enough good sellers, inducing a phase of low search while they exploit the good sellers found. This phase might turn out to be fatal for buyers, which might be thrown out of the market if unable to survive until fresh meat, i.e. a new seller, appears.



(a) Good Sellers Discovery

(b) Quality of Contracts

(c) Uncertain Answers

Fig. 1. Simulation Results in L1 and L2 with 0% of Cheaters for 100 turns

According to our hypothesis, the introduction of reputation in the market has a smoothing effect; in L2 the search seems to hit a constant rate. Considering the uncertain nature of reputation, although both checked and non-checked information is let in, uncertainty reduces (Figure 1c and Figure 2c). There is a clear correlation beetwen good sellers discovery and the uncertainty reduction with reputation.

On the other extreme,in Figure 2a with 90% of agents lying, the L1 and L2 curves for GS are equivalent. The number of good seller discovered is lower than in absence of cheaters, showing how the effect of reputation is attenuated by false information.

The general trend of the market without cheaters is shown in Figure 1b where quality Q is represented against turns of simulation. In both L1 and L2 scenario, the system achieves optimal quality levels during the initial simulation phase; we can only observe a slightly faster L2 convergence. Both in L1 and L2 the process noisily oscillates around the value reached, but in L2 it is smoother, not showing the high peaks as in L1; in L2, the noise derivating by the stock-driven sellers disappearance is smoothed.

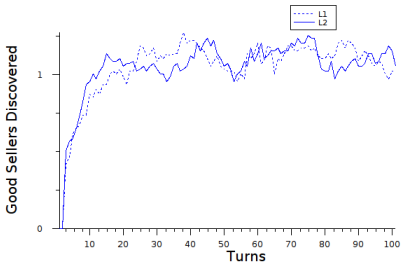
Given the dynamics of GS without liars reaching high peaks and considering the lower number of good sellers discovered when the cheaters are the majority, we would expect a consequent lower quality in a society affected by false information spreading. Instead the effect on quality when informational cheaters are

the 90% of the population, is barely perceivable (compare Figure 2b, Figure 2a and **Table 2**.

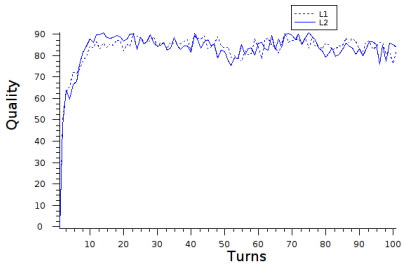
In Figure 1c and 2c trend of uncertainty (represented by answers of *Idontknow* type) is shown for the 100 simulation turns, respectively with only honest agents and with the majority of liars, for both L1 and L2. The curves present a considerably different behavior: in L1, with only image circulating, there is a faster growth at the beginning and then values remain stable on high levels. In L2, the trend of uncertain evaluations starts to decrease after few iterations until reaching a low level. The low level of *Idontknow* answers corresponds to the growth of reputation spreading. So, the spread of reputation appears independent of its truth value, as foreseen from theory. The system is anyway capable to compensate for this large amount of false information, and maintains a comparable quality performance in L1 and L2.

We have the same effect in the two opposite settings to prove the capacity of reputation to reduce uncertainty. The theory in object [2] suggests that the phenomena of uncertainty reduction is an effect emerging not only from more information in L2 than in L1, but a quantitative effect of different types of social evaluation.

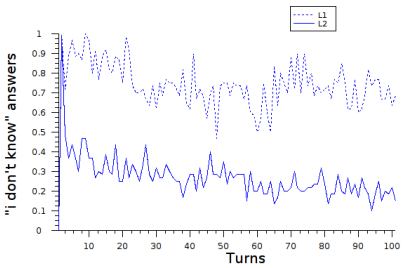
Reputation lets in more information into the system, with a consequent decrease of uncertainty.



(a) Good Sellers Discovery



(b) Quality of Contracts

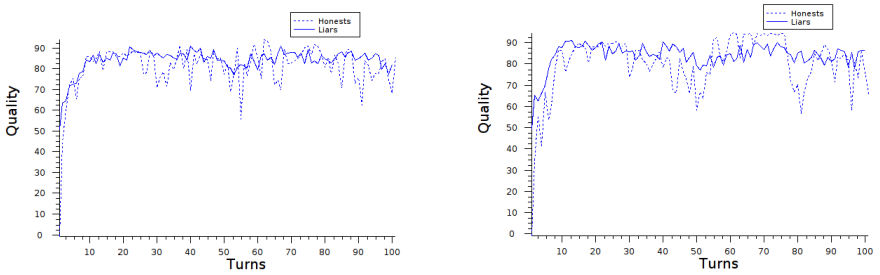


(c) Uncertain Answers

Fig. 2. Simulation Results in L1 and L2 with 90% of Cheaters for 100 turns

For a comparison of honest and cheaters agents, we extract from the simulation results the Good Sellers discovery, Quality and Uncertainty trends of the different strategies (experiments 90CL1 and 90CL2).

Surprisingly the performance of cheaters and honest agents are comparable from all points of view; We report the quality trends in Figure 4a and 4b, for L1 and L2. This hints to a process where cheaters are metabolized in to the system and differences are canceled, a phenomenon that deserves further investigation.



(a) Quality of Contracts in L1 (b) Quality of Contracts in L2

Fig. 3. Simulation Results in L1 and L2 with Cheater against Honest Agents

6 Conclusions

In an uncertain world, humans or intelligent agents, to cope with uncertainty, need to communicate and share information to increase their experiences, and consequently their possibility of success. A model for describing dependencies of different informational domains among agents endowed with different behaviors (honest and liars) was applied to a computer simulation study for partner selection dynamics in a base market. Information exchange is a combined social function in which individuals request, provide, and exchange information with the goal of reducing uncertainty and plays a basic role for groups formation. Agents act as individuals in their intentions but in the Repage market they need to share knowledge for evaluation and this aspect makes agents influence each other in their beliefs formation and revision. Results show that quality levels obtained are comparable in the situation without cheaters and with the majority of cheaters, a condition that shows just a slight quality loss. Only the situation without cheaters and without reputation shows a different trend in good seller discovery, hinting to a different use of information. The good performances of the settings with a high level of cheaters hint to metabolizzation of bad information that will be the object of future studies. False information produces an unexpected impact precisely on the situations in which communication could have been more important.

7 Acknowledgments

This work was supported by the European Community under the FP6 programme (eRep project CIT5-028575). A particular thanks to Rosaria Conte, Jordi Sabater, Isaac Pinyol, Antonietta Di Salvatore, Federica Mattei, Daniela Latorre.

References

1. C. R. Berger and R. J. Calabrese. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research*, 1(2):99–112, 1975.
2. R. Conte and M. Paolucci. *Reputation in Artificial Societies: Social Beliefs for Social Order*. Springer, October 2002.
3. R. Dunbar. The social brain hypothesis. *Evolutionary Anthropology*, 6:178–190, 1998.
4. S. S. Izquierdo and L. R. Izquierdo. The impact of quality uncertainty without asymmetric information on market efficiency. *Journal of Business Research*, 60(8):858–867, August 2007.
5. I. Pinyol, M. Paolucci, J. Sabater-Mir, and R. Conte. Beyond accuracy. reputation for partner selection with lies and retaliation. In *Eighth International Workshop on Multi-Agent-Based Simulation*, pages 134–146, 2007.
6. W. Quattrociocchi and M. Paolucci. Cognition in information evaluation. the effect of reputation in decisions making and learning strategies for discovering good sellers in a base market. In *EUMAS 07 HAMMAMET Tunisia*, 2007.
7. W. Quattrociocchi, M. Paolucci, and R. Conte. Dealing with uncertainty :simulating reputation in an ideal marketplace. In *AAMAS 08 Trust Workshop Cascais Portugal*, 2008.
8. J. Sabater and C. Sierra. Who can you trust: Dealing with deception. In *Proceedings of the Second Workshop on Deception, Fraud and Trust in Agent Societies, Montreal, Canada*, pages 61–69, 2001.
9. J. Sabater-Mir and M. Paolucci. On representation and aggregation of social evaluations in computational trust and reputation models. *Int. J. Approx. Reasoning*, 46(3):458–483, 2007.
10. J. Sabater-Mir, M. Paolucci, and R. Conte. Repage: REPutation and imAGE among limited autonomous partners. *JASSS - Journal of Artificial Societies and Social Simulation*, 9(2), 2006.
11. A. Salvatore, I. Pinyol, M. Paolucci, and J. Sabater-Mir. Grounding reputation experiments. a replication of a simple market with image exchange. In *Third International Model-to-Model Workshop*, pages 32–45, 2007.
12. M. Schillo, P. Funk, and M. Rovatsos. Who can you trust: Dealing with deception. In *Proceedings of the Second Workshop on Deception, Fraud and Trust in Agent Societies, Seattle, USA*, pages 95–106, 1999.

Author Index