W23
Workshop on
Trust in
Agent
Societies
(TRUST)
TRUST IN AGENT SOCIETIES (TRUST-2012)

Organizers:

*Rino FALCONE, Suzanne BARBER, Jordi SABATER-MIR, Munindar SINGH*
DESCRIPTION OF THE WORKSHOP

Trust is important in many kinds of interactions (human-human, human computer, human-human through computers, and more in general among social agents) in order to define those elements (nature, features and interrelations) that are essential in social reliability.

With the growing impact of electronic societies, trust, reputation, privacy, and identity become more and more important. Trust is not just a simple and monolithic concept, it has different facets, levels, and kinds. We can consider: trust in the environment and in the infrastructure (the socio-technical system) including trust in your personal agent and in other mediating agents; trust in the potential partners; trust in the warrantors and authorities (if any). Another growing trend is the use of reputation mechanisms, and in particular the interesting link between trust and reputation. Many computational and theoretical models and approaches to reputation have been developed in the last few years. In all these cases, electronic personas may be created in many different forums (ecommerce, social networks, blogs, etc). Also the identity and associated trustworthiness must be ascertained for reliable interactions and transactions.

Trust is foundational to the notion of agency and for its defining relation of acting "on behalf of". It is also critical for modeling and supporting groups and teams, organizations, co-ordination, negotiation, with the related trade-off between individual utility and collective interest. Further, computer technology can even break trust relationships already held in human organizations and relations, and favor additional problems of deception and trust.

The aim of the workshop is to bring together researchers (even from different disciplines) who can contribute to a better understanding of trust and reputation in agent societies. Most agent models assume trustworthy communication to exist between agents. However, this ideal situation is seldom met in reality. In the human societies, many techniques (e.g. contracts, signatures, long-term personal
relationships, reputation) have been evolved over time to detect and prevent deception and fraud in communication, exchanges and relations, and hence to assure trust between agents. Artificial societies will need analogous techniques.

We encourage an interdisciplinary focus of the workshop - although focused on virtual environments and artificial agents - as well as presentations of a wide range of models of deception, fraud, reputation and trust building.

Suitable submissions may describe the key elements of social reliability and any topics closely associated with trust such as reputation, privacy, norms, and identity. All facets of trust are relevant. These include trust in • the environment and the infrastructure (as in a socio-technical system) • user assisting and mediating agents; • potential partners, including their electronic personas in commerce and social media; • warrantors and authorities.

We welcome computational and theoretical models and approaches to trust as well as applications and empirical studies on trust. We particularly encourage interdisciplinary contributions that shed new light on the above topics.

This edition of the workshop will emphasize the theme of “Trust and Agreement”.

The topics of interest include, but are not restricted to, applications, concepts, models, theories, mechanisms (including architecture, design, and protocols), techniques, and evaluations of

• Trust
• Deception and fraud and its detection and prevention
• Reputation
• Privacy and access control
• Identity in virtual worlds
• Autonomy, delegation, ownership
• Policies, interoperability, protocols, ontologies, and standards
• Scalability and distribution
• Test-beds and frameworks
• Legal aspects
• Organizations and institutions, include regulation and regimentation
• SPECIAL THEME: Trust and Agreement
WORKSHOP ORGANIZERS

Rino FALCONE - ISTC-CNR – Italy, rino.falcone@istc.cnr.it;
Suzanne BARBER - The University of Texas – USA;
Jordi SABATER-MIR - IIIA-CSIC – Spain;
Munindar SINGH - North Carolina State University – USA

PROGRAM COMMITTEE:

▪ Suzanne BARBER - Computer Engineering, The University of Texas, USA
▪ Cristiano CASTELFRANCHI - Cognitive Science, ISTC National Research Council, Italy
▪ Robert DEMOLOMBE - Computer Science, Institut de Recherche en Informatique, Toulouse, France
▪ Torsten EYMANN - Department of Information Systems, University of Bayreuth
▪ Rino FALCONE - Cognitive Science, ISTC National Research Council Italy
▪ Andrew JONES - Department of Computer Science, King's College London, U.K.
▪ Catholijn JONKER - Computer Science, Vrije Universiteit Amsterdam, The Netherlands
▪ Yung-Ming LI - Computer Science, National Chiao Tung University, Taiwan
▪ Churn-Jung LIAU - Institute of Information Science, Academia Sinica, Taiwan
▪ Emiliano LORINI - Computer Science, IRIT, France
▪ Stephen MARSH - Computer Science, National Research Council, Canada
▪ Yuko MURAYAMA - Computer Science, Iwate Prefectural University, JAPAN
▪ Mario PAOLUCCI - Computer Science, ISTC National Research Council, Italy
▪ Jordi SABATER-MIR - Computer Science, IIIA-CSIC, Spain
▪ Sandip SEN - Computer Science, University of Tulsa - USA
▪ Munindar SINGH - Computer Science - North Carolina State University, USA
▪ Chris SNIJDERS – Sociology - Eindhoven University of Technology, The Netherlands
▪ Eugene STAAB - Computer Science – IMC-AG, Germany
▪ Syed Waqar JAFFRY - Computer Science - VU University Amsterdam, The Netherlands
Reputation Estimation based on Online Social Network Structure:  
A Relational Capital Model  
Suratna Budalakoti, K. Suzanne Barber  

Reputation-based Trust Evaluations through Diversity  
Anthony Etuk, Timothy J. Norman, and Murat Sensoy  

A Subjectivity Alignment Approach for Effective Reputation Computation  
Hui Fang, Jie Zhang, Murat Sensoy, and Nadia Magnenat Thalmann  

A Model for Augmenting Trust Management using Argumentation  
Sharmin Jalal, Karl N. Levitt, Jeff Rowe, Elizabeth Sklar, and Simon Parsons  

A Formal Argumentation Dialogue for Personalised Trust Communication  
Andrew Koster, Jordi Sabater-Mir, and Marco Schorlemmer  

Reasoning about Advisors for Seller Selection in E-Marketplaces via POMDPs  
Frans A. Oliehoek, Ashwini A. Gokhale, and Jie Zhang  

Trusting the Messenger and the Message  
WORKSHOP SCHEDULE

MORNING
Accepted Papers presentation

AFTERNOON
Invited Talks and Discussion

Invited Talks:

Stephen Cranefield: “The role of trust in virtual worlds”

Frank Dignum: “Can I trust my teammate if she is an agent (or a robot)?”

Mario Paolucci: “Trust in the system, trust in peers: Agent-based models of Peer review in Science”

Van Parunak: “From Reliability to Trustworthiness in Complex Autonomous Systems”
Reputation Estimation based on Online Social Network Structure: A Relational Capital Model

Suratna Budalakoti\textsuperscript{1} and K. Suzanne Barber\textsuperscript{2}

\textsuperscript{1}suratna@utexas.edu
\textsuperscript{2}sbarber@identity.utexas.edu
The University of Texas at Austin

\textbf{Abstract.} The problem of identifying influencers in an online social network (OSN) has received a lot of attention recently. However, the influence a user has on an OSN depends a lot on how active they are, besides their reputation, or the quality of information they are capable of providing. This can be misleading; a common observation about many OSNs is that the most influential users there are not well-known in the real world. In contrast, we investigate the problem of automatically identifying reputed users on an OSN, independent of their level of activity. The investigation is based on a social/relational capital generation based model of user behavior on OSNs, where users pay each other in social capital to make connections. We formalize this idea using a tournament-based model known as fair bets. The model is tested on simulated OSN data, where it proves to be more accurate than baseline approaches for reputation aggregation in OSNs, such as PageRank.

\textbf{Keywords:} Trust, Reputation, Social Capital, Social Networks

1 Introduction

A remarkable development of the last few years has been the move online, from anonymity or pseudo-anonymity, to identities that can be mapped to people in the real world. This trend, combined with the increasing reliance on the Internet for information, means that an individual’s online reputation is becoming an increasingly important factor in how they are perceived by others offline. There are many problems with using online information to draw conclusions, as it is an imperfect record of real world activity. However, given its easy availability, it is inevitable that online information will be used more and more to draw conclusions about an individual’s real world profile.

An interesting question then is, what can be said about an individual’s real-world reputation, given online information about them. We focus on a narrower version of the problem: what can a member’s position in the structure of an online social network (OSN) say about her real world trust networks, and her reputation there. The term reputation is defined here, following Jøsang et al. [1], as “a collective measure of trustworthiness, based on the referrals or ratings of members in a community”. The community we are interested in is an individual’s professional peer group,
or the community of people interested in the topics she claims competence in. This peer community is (imperfectly) reflected in OSN data, in the form of other member profiles on the OSN. Similarly, real world peer trust ratings are often unavailable. However, the behavior of an individual’s peers on an OSN, when they interact with her, can provide clues about how trusted she is. Falcone and Castelfranchi [2] note that “to be trusted: i) increases the chance to be requested or accepted as a partner for exchange or cooperation; ii) improves the price, the contract that the agent can obtain”. Looking for these signals in OSN data can help identify which members have a high reputation in the real world.

We focus on reputation in professional settings, as computational models of personal trust and reputation raise philosophically complicated issues [3].

The related problem of identifying high-profile online users, or influencers, has received significant attention [4, 5] in the OSN research community. Usually, influence is measured as the extent to which a member’s behavior affects her peers on an OSN, for example, by convincing them to buy a product, or be favorably disposed towards a particular opinion. Such information can be very useful from a marketing perspective. Also, it is often assumed that a person who is highly reputed in a topic would have higher influence than a less reputed one. Thus, for example, the list of topic-sensitive influencers [6], or extremely influential people in a topic on an OSN, would ideally consist of individuals known to be experts on the topic. However, this is not always the case. For example, Khrabov et al. [7] found that many influencers on the popular microblogging website Twitter³ are relatively unknown in the real world. A major reason for this anomaly is the extremely competitive nature of the quest for online influence, in terms of time and effort. Many people who are highly reputed in their fields may not be willing to invest these to the same extent as less reputed people.

This is not a concern for many applications. For example, if marketeers need to know which users on an OSN should be engaged to market a product, they largely care about their online influence, not their real world reputation. On the other hand, there are many scenarios where reputation might be the primary concern. For example, many people rely on information shared on blogs, or microblogging forums (e.g., Twitter) to find information in their areas of interest. It is likely many of them would be at least as interested in highly reputed users, as in influential ones. Or if a company is looking to fill a job position⁴, reputation might be a greater concern than influence. Also in the marketing domain, Carl [8] has argued that an overwhelming majority of word-of-mouth marketing takes places offline, in which case reputation is a factor worth considering. Separating reputation and influence will also provide better insight into the relationship between online and real world prestige.

³ www.twitter.com
⁴ A number of websites, for example, LinkedIn(www.linkedin.com), and StackOverflow (www.stackoverflow.com), combine professional social networking or expert finding with recruitment solutions.
We base our reputation estimation algorithms on the concepts of social and relational capital [9, 2]. The intuition is that users on an OSN that have a high reputation have more relational capital, that is a larger number of reputed people are positively disposed towards them, as compared to less reputed users. As a result they need to expend less effort to attain the same level of connections (in number and quality), as less reputed users. This can be modeled as follows: assuming users on an OSN paid each other in some form of capital to accept social connections. Then highly reputed users are those that do not have to pay very often, but themselves get paid quite often. In other words, when they pay, on average, they can afford to pay much more. This idea is formalized in later sections using the fair bets model [10, 11].

2 Background and Related Work

A natural representation of an OSN is as a graph, with members represented as vertices, and edges representing an interaction. The edges may be directed or undirected depending on the interaction type: a symmetric connection might be represented by an undirected edge, while a directed edge can be used to signify who sent the invitation. Directed edges can also be used for OSNs that allow for asymmetric connection (a connection need not be accepted to exist, as, for example, in Twitter). A common assumption [12, 13] is that, if user \(A\) took the initiative to send an invitation to user \(B\), then that counts as an endorsement from user \(A\) to \(B\).

An interesting aspect of this scenario is that the interactions that result in online connections take place offline. As a result, it is usually impossible to know what prompted user \(A\) to connect to user \(B\), how often they have interacted in the past, and what \(A\)'s experiences were during these interactions. It is also not known that if, on the other hand, user \(A\) has not invited user \(B\), it is because they are not acquainted with each other or have not interacted sufficiently, or because user \(A\) distrusts user \(B\).

As a result, many traditional reputation aggregation methods, such as belief models [14] and reinforcement learning-based methods [15], do not fit naturally in this scenario. On the other hand, due to the graph-based structure of problem representation, eigenvector-based approaches [16, 17] inspired by PageRank [18] are a better fit for the problem. These approaches posit a recursive definition of a vertex’s reputation score in the graph, as a weighted sum of the reputation score of all the vertices that point to it.

Variations of eigenvector-based approaches are also popular in the OSN community for the task of influence estimation [6]. While well-suited to finding influencers, in their present form, they have significant drawbacks when it comes to reputation estimation. They do not take into account a user’s activity level when estimating their score. This is a reasonable decision when identifying influencers: if an influential user is highly active on an OSN, this does not detract from the level of influence they have. Reputation, however, has a different dynamic.

A user’s influence on an OSN depends on two factors: a) her reputation, which determines the value other people see in a connection with her,
and b) her visibility on the OSN, that is, the likelihood that she will be noticed by other users. A user with a lower reputation level can become more influential by increasing her visibility. This can be achieved through increased activity (for example, by engaging other members in different ways), or increased connectivity (by sending more invitations). On the other hand, users that are highly reputed in the real world, are likely to receive invitations even with lower level of online activity. In other words, a user’s reputation in the real world can be roughly measured as the degree of influence she has on an OSN, normalized by her level of activity. In the next section, we formalize this intuition using the concept of social capital.

3 Social and Relational Capital in an OSN

Social capital [9, 19] is a concept used to refer to value generated by the resources that can be mobilized via an individual’s social connections. Boxman et al. [20] defined social capital as “The number of people who can be expected to provide support and the resources those people have at their disposal”. However, in the past, the term has been used ambiguously to also refer to the collective advantage that a group attains by being better connected. Due to this reason, Falcone and Castelfranchi have proposed the term relational capital be used instead, to represent “how much the agent is valued by other agents in a given market for a given task” [2], with the intent of separating an individual’s advantage, from that of the entire group.

Putnam [9] divides social capital into bonding and bridging social capital. Bonding social (or relational) capital is the capital generated for an individual by the connections made within a social community, while bridging social capital refers to the capital an individual generates by links made across social communities. From this perspective, OSNs can be seen as being used by its members for the purpose of maintaining and growing their social capital [19]. Thus for example, when users connect with old acquaintances on an OSN, this helps to maintain or increase the existing bonding capital, which can be mobilized later if required. On the other hand, connections made outside one’s community, for example to highly reputed or influential people in one’s professional field, increase one’s bridging social capital.

These two types of social capital, bridging and bonding, can be mapped to two different types of network growth patterns documented on online networks. The fitness-based preferential attachment model [21], where the in-degree of each vertex grows with a probability proportional to its fitness (which can correspond to reputation), and its current in-degree, can be seen as a result of users trying to build their bridging capital. Bonding can be mapped to the phenomenon on triadic closure [22, 23] observed on OSNs, where networks grow by closing incomplete triangles (connection of a connection).

A few conclusions can be drawn based on this interpretation. People are more eager to connect to highly reputed users as they have more bridging capital, and will work harder to discover them on the OSN, looking beyond their immediate network. As a result, users who are highly reputed
will receive more and stronger incoming links, even with low visibility on the OSN. Also, the overall quality of links created via triadic closure will be higher, as the peer group of users with higher reputation is likely to have higher than average authority, assuming the concept of homophily [24] holds.

On the other hand, users with lower offline reputation might also prove themselves valuable users to connect to, for the purpose of generating relational capital, but this is likely to be via one of two ways: a) achieving high visibility on the OSN via higher levels of activity, or b) generating relational capital for themselves by connecting to others who already have that capital (due to their higher reputations).

In the next section, we discuss how these ideas can be formalized in a mathematical model, called the fair bets model.

4 The Fair Bets Model

The fair bets model has traditionally been used [25, 10, 11] to rank players in round-robin tournaments. A player’s behavior in the model is visualized as follows: she is allowed to bet a certain amount of money per game, which has to be fixed across all games she plays, irrespective of the opponent. She forfeits this amount to her opponent if she loses the game, and if she wins, she is awarded the amount bet by her opponent. Then the amount she can afford to bet per game is the score assigned to her.

Mathematically, this can be represented as follows: construct a graph $G$, with each player $i$ as a vertex $v_i$, and assume that each player has played at most one game against any other player (this can be generalized). Then draw an edge directed from the loser of the game to the winner. A matrix representation of this graph can then be written as $V$, where $v_{ij} = 1$ if there is an edge directed from player $i$ to $j$ in the graph. Then the fair bets score $a_j$ of player $j$ follows the following equation:

$$\sum_{i=1}^{N} v_{ij} a_i = \sum_{i=1}^{N} v_{ji} a_j$$

(1)

That is, a player $j$’s payout per game ($a_j$) is the amount she makes in total, divided by the number of games lost. In matrix form:

$$V^T a = C a$$

(2)

$C$ is a diagonal matrix with $C_{ii}$ equal to the sum of the $i$th row of $V$, i.e., $C_{ii} = \sum_{k=1}^{N} v_{ik}$.

A relationship can be established between PageRank and fair bets scores [25]. For a right stochastic matrix $P$, the PageRank vector $r$ corresponding to the matrix is given by the equation:

$$P^T r = r$$

(3)

Let $P = C^{-1}V$. That is, $P$ is the stochastic matrix given by normalizing all the rows of some matrix $V$ to 1.
Then, equation (3) can be written as:

\[ V^T C^{-1} r = r \]  (4)

\[ \Rightarrow C^{-1} V^T (C^{-1} r) = C^{-1} r \]  (5)

Setting \( a = C^{-1} r \) gives:

\[ C^{-1} V^T a = a \]  (6)

\[ \Rightarrow V^T a = Ca \]  (7)

which is the same as equation (2). That is, if the PageRank vector for a stochastic matrix \( P \) is given by \( r \), then the fair bets vector of the original graph matrix \( V = CP \) is given by \( a = C^{-1} r \). Thus, mathematically, the fair bets score of a vertex in a graph is equal to its PageRank score, divided by its outdegree.

### 4.1 Fair Bets as Relational Capital

In the context of online social networks, fair bets can be viewed as a model of relational capital accumulation and expenditure. An OSN’s members tend to connect to members who are likely to provide useful information or members who can help them in achieving professional growth, that is, members with high relational capital. This initial expenditure of relational capital spent in sending invitations pays off, as the user accumulates invitations (payments) in return. Highly reputed users receive multiple invitations without making a significant effort, while the payoff for less reputed users is lower. On the other hand, if highly reputed users decide to spend their capital by sending invitations, their payoff per invitation is proportionately much larger.

### 4.2 Variations on The Fair Bets Model

For an OSN graph, the standard fair bets model assumes a linear relationship between a vertex’s authority score and its outdegree. As the fair bets score is given as the PageRank score divided by the outdegree, a user’s PageRank must grow at a constant rate with invitations sent. Whether such a relationship holds, depends on the dynamics on the network. In practise, a user’s network is likely to saturate over time, so that new invitations sent lead to fewer and fewer invitations back.

Consider an OSN with the following dynamic: new users that join the network see an initial spurt in the rate at which their network grows. Each invitation that they send out increases the circle to which they are visible, leading to many more invitations back to them. Over time, though, as most connections-of-connections are now their first degree connections, they receive fewer invitations back for each invitation sent. This dynamic can be modeled by assuming that the expected value of links received once \( k \) invites have been sent is \( \frac{1}{k} \). This expected value includes both the probability of receiving a link, and the value of the link. Then, since \( \sum_{i=1}^{k} \frac{1}{i} \) can be approximated as \( \log k \). Given this kind of dynamic, a better measure of a user’s reputation, might be:
where $\pi_i$ is the user’s PageRank score, and $o_i$ is her outdegree, or the number of invitations sent. This measure can be referred to as the log fair bets. Log fair bets can be interpreted as assuming that the arrival patterns of incoming links follows a power law distribution with respect to time (measured by outdegree). On the other hand, assuming that the mean value of an incoming link falls even more steeply with activity, say the expected value of an incoming link for a user is $\frac{1}{k^2}$. Since $H_{k,m} = \sum_{i=1}^{k} i^m$ is the definition of a generalized Harmonic number, we can define a new measure, the Harmonic fair bets measure $h_i$ for user $i$, set as:

$$h_i = \frac{\pi_i}{H(o_i, m)}$$

In our experiments, $m$ was set as 2. In the next section, we investigate, using a simulation based approach, various variants of the fair bets measure, including log fair bets, and their suitability for different types of OSNs, based on its characteristics.

5 Simulation Experiments

We consider two models of social network growth based on past research: preferential attachment models, and triadic closure based models.
Preferential Attachment Model: As mentioned earlier, this model can be seen as describing OSNs where bridging social capital dominates. In such OSNs, users tend to seek out experts in their area, even when they are not part of their immediate social circle. However, they are still more likely to find users who have lots of connections, as opposed to users who have fewer connections.

Triadic Closure Model: This model is better suited to describe the growth dynamics of OSNs where users are looking to maintain their bonding social capital. Such networks grow by closing ‘triangles’. Users grow their network by inviting their second degree connections to become their first degree connections.

The preferential attachment (PA) model is simulated as follows: we start with a graph of zero vertices. At each timestep, with a small probability (0.02), a new vertex is added to the graph. The vertex is assigned one of five fitness levels (1–5). Level 5 corresponds to a user with the highest reputation level, while level 1 corresponds to one with the lowest. The levels are drawn from a power law distribution, so that the probability of drawing a level 5 edge is one-fifth that of drawing a level 1 vertex. If a new vertex is not drawn, one of the existing vertices $v_1$ sends an invitation to connect to another vertex $v_2$, where the probability of selecting any vertex $v_2$ is proportional to the product of its indegree and its fitness level. The intuition is that the probability that $v_1$ will see $v_2$’s profile is proportional to $v_2$’s popularity, which is its indegree, while the probability that $v_1$ will want to connect to the profile is proportional to its fitness level. The model is similar to the one described in by Bianconi et al. [21].

The triadic closure (TC) model is set up to draw vertices from a power law distribution with 5 levels, similar to the PA model. A vertex $v_1$ in the TC model grows its network as follows: at each timestep it selects one of its first-degree connections, say $v_2$, at random. From the list of connections of $v_2$, it selects a new connection $v_3$ to invite to connect. The probability of selection of a particular $v_3$ is proportional to its fitness level. The model is based on the one described by Jin et al. [23], with the addition of the fitness level concept, to model reputation.

Table 1 shows the rank correlation, using Spearman’s $\rho$ measure, of the assigned fitness level of a user, to the PageRank, Fair Bets (FB), and Harmonic Fair Bets (HFB) score assigned to them, for the two growth models. As can be seen both FB and HFB correlate much better with the fitness level, compared to PageRank. This shows that PageRank is not a...
good measure of reputation levels. In contrast, Fair Bets and Harmonic Fair Bets do much better. Figure 1 shows how the three measures vary with user activity levels for all Level 2 vertices, for the PA model. The vertices on the x-axis are ordered chronologically, so the ones on the left have had the most time to be active, while those on the right have had the least. As all vertices shown are of Level 2, ideally they should have the same reputation score. The figure shows the deviation from this ideal score, that is, the error around 0. As can be seen, PageRank is very sensitive to activity. The longer the vertex has been in the system, the higher its score. Fair Bets, on the other hand, seems to over-correct for activity level, with less active vertices receiving a disproportionately high score. Harmonic Fair Bets sticks closest to the 0 line. However, Fair Bets performs better for the triadic closure model (Table 1). The reasons for this are not obvious, given the complexity of even seemingly simple graph evolution models. We plan to investigate the reasons for this further in future work.

6 Conclusion

This paper introduces a new problem to the field of reputation modeling. Given the structure of an online social network graph, what can we say about their offline reputation. We propose a possible approach to addressing this based on the following observation: highly reputed individuals have a lot of social/relational capital, which causes them to be treated differently from individuals with a relatively lower reputation. Taking this observation into account, it is possible to estimate user reputation, even when explicit trust values are not available. Based on this, we develop an approach for offline reputation estimation in OSN, and test it on simulations based on standard models of OSN growth. We show that the PageRank measure, while suitable for measuring influence, is not effective for estimating reputation values, as it is too easily mislead by activity levels. On the other hand, two new reputation measures, fair bets and harmonic fair bets, give more promising results.

References


Reputation-based Trust Evaluations through Diversity

Anthony Etuk, Timothy J. Norman, and Murat Şensoy
Department of Computing Science, University of Aberdeen, Scotland, UK
{aetuk,t.j.norman,m.sensoy}@abdn.ac.uk

Abstract. Trust and reputation are significant components in open dynamic systems for making informed and reliable decisions. State-of-the-art trust models that exploit reputational evidence generally rely on reports from as many sources as possible. Situations exist, however, where seeking evidence from all possible sources is unrealistic. This is particularly the case in resource-constrained environments where querying information sources is costly, for instance in terms of time and bandwidth. This paper describes an approach that exploits diversity among information sources in order to select a small number of candidates to query for reputational evidence. We demonstrate that reliable decisions can be reached using evidence from small groups of individuals. We show that our approach is robust in contexts of variable trust in reputational sources and to a degree of deception.

Categories and Subject Descriptors: I.2.11 [Distributed Artificial Intelligence]: Multi-Agent Systems
General Terms: Experimentation, Performance
Keywords: Diversity, Reputation, Trust

1 Introduction

Reputation-based trust is a critical mechanism in large, open, and dynamic systems, where agents must interact with one another in order to achieve their goals. Agents operating in such environments often rely on indirect experience acquired from their peers in order to make informed decisions, especially when direct experience on a subject is lacking or insufficient [10]. Whereas state-of-the-art trust models exploiting this kind of evidence generally rely on reports from as many sources as possible, in the physical world capturing and distributing evidence can be costly. For instance, in distributed environments such as peer-to-peer networks, sensor networks, and pervasive computing, each participant is responsible for collecting and combining evidence from others due to lack of central authority or repository. A major constraint in such systems is bandwidth, motivating the need to minimise the number of messages exchanged in order to arrive at a decision. As a result, reputation assessments are often based on a subset of evidence, usually from the agent’s neighbourhood [3], an approach which in itself can be problematic, as it does not make use of all the information available, and therefore is prone to biases and deception.
Motivated by this problem, we present an approach for minimising the costs associated with making effective trust assessments, while still remaining robust to biases and deception. In particular, we exploit diversity among information sources to intelligently sample from the crowd of reputation sources. Our notion of diversity is inspired by the work of Surowiecki et al. [8]. Their work highlights some interesting criteria for effective decisions in large groups of individuals. Diversity ensures that the experiences of different agents based on their private information is exploited in a decision process. As agents most often operate in social contexts where interaction with others is inevitable, so much could go to influence their behaviour. Agent behaviour could be conditioned by organisation, profession, or proximity to other agents in a network. While Surowiecki et al. consider such effects detrimental to a decision process because of the possibilities of collusive behaviour and subjectivity of opinions, we see great potential. For example, where logical subgroups exist in the agent population as informed by a feature-behaviour correlation, we can exploit this to limit the number of agents queried for evidence. In the sections that follow, we demonstrate how this concept can be employed in a manner that leads to positive outcomes.

The rest of this paper is organised as follows: Section 2 highlights the different components of the diversity model. An evaluation of the approach is presented in Section 3, and in Section 4 we conclude with a discussion and avenues for future research.

2 The Diversity Model

The Diversity Model (DM) enables an agent to arrive at reliable decisions using evidence from small groups of individuals. The model employs trust and machine learning techniques in order to build models of information sources from which potential candidates may be sampled for evidence.

We consider an evaluator $x$, who wishes to evaluate the truth of a proposition $\rho$, and has access to a set of information sources $A$, where individual information sources or agents are denoted $x, y, \ldots \in A$. The notation $x$ will be used in the course of this paper to denote an agent acting in the capacity of an evaluator, while the notation $y$ will be used to represent an agent regarded as an information source. In a more general sense, an agent is regarded as a tuple $(ID, F, R)$, where $ID$ denotes a unique identifier, $F$ is a set of features, and $R$ is a set of past reports.

A report $R$, is an opinion about a subject $\rho$, provided by an agent $y$, to an evaluator $x$, in response to a query. An agent $y$, records its perceived opinion about $\rho$ as $R_{y,\rho}$, and reports $R_{y-x,\rho}$ when queried by $x$. The variable $t$ denotes a time step associated with a report from $y$, such that $R_{y,\rho}^t$ represents a report at time $t$. Consequently, $R_{y-x,\rho}^{t+k}$ denotes the set of reports received by $x$ from $y$ between the interval $t$ and $t+k$.

Let $F$ denote a finite set of features, such that $f_1, f_2, \ldots, f_d \in F$. We define a feature as an observable attribute of an agent, e.g., an agent’s organisation or its location. An evaluator $x$, has a view on the relative importance of features
represented by the vector \( \langle w^x_1, w^x_2, \ldots, w^x_d \rangle \), where \( w^x_i \) is \( x \)'s view of the importance of feature \( f_i \), and \( w^x_i : [0,1] \). Subsequently, an evaluator uses this metric to compute the similarity between different agents. Similarity between agents is, therefore, a measure of the distance between their features as informed by the vector of weights on the features. We employ the weighted Euclidean distance to compute similarity between any two agents \( y_1 \) and \( y_2 \) as:

\[
D^F_{y_1,y_2} = \|y_1 - y_2\| \cdot \mathcal{F} = \sum_{k=1}^{[F]} w_k \left( f_k y_1 - f_k y_2 \right)^2.
\]  

(1)

2.1 Group

Let \( G \) denote a stratification on \( A \), such that \( G = \{G_i, G_j, \ldots, G_m\} \) where \( G_i \cap G_j = \emptyset \), if \( i \neq j \). We define a group \( G_i \) as a collection of homogenous agents, such that \( \{x : x \in G_i, G_i \in G\} = A \). Groups are formed subjectively by an agent who attempts to disambiguate what metrics lead to a better stratification of information sources. The group formation process is discussed in Section 2.3. However, the aim of an agent in partitioning the population, is to provide a suitable generalisation of information sources using different distinguishing characteristics. Subsequently, an agent exploits this model to limit the number of sources queried for evidence, and to protect itself against deception. Agents are partitioned into groups based on how similar they are to one another, as specified by a similarity metric. We denote by \( G_i(y) \), an agent \( y \)'s membership of a group \( G_i \). An agent \( x \), maintains two parameters \( \delta^F_{G_i} \) and \( \delta^B_{G_i} \), which denote the feature-based similarity and the behaviour-based similarity of a group respectively.

The feature similarity \( \delta^F_{G_i} \) of a group is the degree of similarity of members of the group given their features. This parameter is measured by computing the average weighted distance between pairs of agents in the group as follows:

\[
\delta^F_{G_i} = 1 - \frac{2}{n(n-1)} \sum_{y_p,y_q \in G_i} D^F_{y_p,y_q} \text{ where } p < q, \text{ and } n = |G_i|. 
\]  

(2)

An evaluator learns over time the importance of different features while computing similarity. Consider for instance, the following feature set \( \langle \text{age, profession, location} \rangle \) describing agents in a population. An agent may assign different weights to different features while measuring similarity. For example, an agent could measure similarity using \( \text{age} \), or \( \text{location} \), a combination of \( \text{age} \) and \( \text{location} \), etc. Although different feature subsets may define different subgroups in the population, not all feature subsets might be distinguishing enough for identifying relevant subgroups in the population. In an example scenario, an agent wishing to evaluate the reliability of a delivery company, may learn informing subgroups in the population of reputation providers, by partitioning agents based on their location for instance, rather than their age or profession. The fact that some locations may be easily accessible (e.g., metropolitan areas), than others (e.g., rural areas), might impact on the satisfaction level of agents obtaining services from.
the company, and potentially reveal a relationship between the feature location and the ratings obtained from agents.

The behaviour similarity of a group \( \delta^{x}_{G_i} \), is a subjective measure of the likelihood of agents in the group behaving in a similar manner. In the context of this paper, behaviour is represented by the report of agents. It is important to emphasise here that behaviour similarity does not capture a semblance of the agents based on their level of trustworthiness (e.g. honest, deceitful), rather it is a measure of the consistency of agents in giving similar reports (be it honest or deceptive ones), in response to the same query. Although agents belonging to the same group may be regarded as having the same level of trustworthiness as depicted in Section 2.2, in our model this condition alone does not satisfy the criteria for grouping agents. It is possible for dissimilar agents to have similar level of trustworthiness (e.g. agents from different but highly reputable organisations). In order to effectively exploit diversity in the system, our model requires agents in a group to be similar both in feature and behaviour. To compute the \( \delta^{x}_{G_i} \) of a group, a report matrix is constructed as illustrated in Figure 1. The

\[
\begin{array}{c|cccc}
  t_1 & t_2 & t_3 & t_4 \\
  \hline
  y_1 & 1 & 5 & 1 & 4 \\
  y_2 & 4 & 1 & 5 & 1 \\
  y_3 & 1 & 5 & 2 & 4 \\
\end{array}
\]

Fig. 1. Report matrix for similarity calculation

matrix captures the rating provided by different agents in a given sampling interval. A sampling interval is the time frame for which reports from different agents are considered, and is the same for all the agents. In Figure 1 for example, the sampling interval considered is \( t_1 : t_4 \) (i.e. \( t_1, t_2, t_3, t_4 \)), and the reports from the agents could be represented as \( R_{t_1 : t_4}^{i} \), \( i = 1, \ldots, 3 \). Also, agents \( y_1 \) and \( y_3 \) with report vectors \( \langle 1, 5, 1, 4 \rangle \) and \( \langle 1, 5, 2, 4 \rangle \) respectively, may be considered much more similar to each other than agents \( y_1 \) and \( y_2 \) with report vectors \( \langle 1, 5, 1, 4 \rangle \) and \( \langle 4, 1, 5, 1 \rangle \) respectively. Details for the computation of this measure is given in Equation 3 and Equation 4.

\[
D^{R}_{y_1, y_2} = \frac{1}{h} \sqrt{\sum_{\tau \in H} \left( R_{y_1, \rho}^{\tau} - R_{y_2, \rho}^{\tau} \right)^2},
\]  

(3)

where \( h \in H \) represents the number of past reports taken into consideration. Following Equation 3, \( \delta^{x}_{G_i} \) of a group \( G_i \) can be computed as:

\[
\delta^{x}_{G_i} = 1 - \frac{2}{n(n-1)} \sum_{y_p, y_q \in G_i} D^{R}_{y_p, y_q} \text{ where } p < q, \text{ and } n = |G_i|.
\]  

(4)

2.2 Group Trust and Subjective Logic

An evaluator depending on evidence from third party sources faces the risk of misleading reports from these sources. Not all agents may act in a benevolent
manner or even possess a required level of expertise to report on a subject. Sometimes information sources may exaggerate perceived opinion, or offer testimonies that are outrightly false. Finding ways to reduce the influence of misleading reports from third-party sources is a fundamental problem in reputation systems [10]. One way of mitigating against this problem, is maintaining a reputation of the information sources, and using this to determine the weight given to their reports [9].

Subjective Logic (SL) [1] is a belief calculus which allows agents to express opinions as degrees of belief, disbelief, and uncertainty about propositions. Binary propositions, such as agent y, is trustworthy concerning r, lead to opinions which are equivalent to a beta distribution. SL contains operations to represent consensus, recommendation, and ordering, which are useful tools for evidence aggregation. We adopt SL to represent trust because it provides an intuitive way to support such belief. An evaluator’s opinion about another entity can then be used by an evaluator to represent the belief an entity has in another, and a way to aggregate evidence to support such belief. An evaluator’s opinion about an agent y, reporting correctly on r is represented in Subjective Logic as a tuple:

\[
\omega_{xy} = (b_{xy}, d_{xy}, u_{xy}, a_{xy})
\]

where \(b_{xy} + d_{xy} + u_{xy} = 1\), and \(b_{xy}, d_{xy}, u_{xy}, a_{xy} \in [0, 1]\). (5)

In the above representation, \(b_{xy}\), \(d_{xy}\), \(u_{xy}\), \(a_{xy}\) represent the degrees of belief, disbelief, uncertainty, and the base rate (a priori probability in the absence of evidence) respectively. Opinions are formed on the basis of positive and negative evidence. The variables \(r_{xy}\) and \(s_{xy}\), represent the number of positive and negative past observations of x about y respectively, and can be used by x to obtain an opinion about y as follows:

\[
b_{xy} = \frac{r_{xy}}{r_{xy} + s_{xy} + 2}, d_{xy} = \frac{s_{xy}}{r_{xy} + s_{xy} + 2}, u_{xy} = \frac{a_{xy}}{r_{xy} + s_{xy} + 2}
\]

An opinion’s probability expectation value computed using Equation 6, can be used by x as a measure of y’s trustworthiness with respect to r.

\[
\tau_{xy} = b_{xy} + a_{xy} \times u_{xy} = \frac{r_{xy} + a_{xy}}{r_{xy} + s_{xy} + 2}
\]

The base rate parameter \(a_{xy}\), also known as the relative atomicity, represents a priori degree of trust r has about y giving accurate report about r before any evidence has been received. The parameter determines how uncertainty shall contribute to the computed expectation value. The default value of \(a_{xy}\) is 0.5 [2], which means that before any positive or negative evidence has been received, both outcomes are considered equally likely.

The trust value for a group \(G_i\) is based on past interactions with members of the group, and computed as a function of the trust of the individual members encountered from the group.

\[
\tau_{G_i} = \frac{\sum_{y \in G_i} \tau_{xy}}{|G_i|}.
\]
2.3 Learning Diversity

We define Diversity as a function $\Delta : 2^A \rightarrow G$ that maps the feature set and past reports (behaviour) of agents to a set of groups. We take as a working assumption, that there may be some correlation between the features of an agent and its behaviour. Where this exists, we could exploit information from observable features of agents, as well as evidence from their past behaviour to build a model of diversity. Diversity learning may be carried out in two stages: the first stage involves an attempt at disambiguating what metrics lead to a better stratification of the population of agents. The best metric in our estimation is one that produces the highest feature-behaviour correlation, such that the probability of agents in the same group giving similar reports is maximized. We refer to this correlation as group behaviour. In the second stage, the learned metric is employed to partition agents into semi-homogenous subgroups. We focus here on the process of group formation by assuming a learned metric.\footnote{Relaxing this assumption is left to a future work.} We employ a clustering mechanism that incorporates a feature threshold $fT$, and a behaviour threshold $bT$ in order to control the formation of clusters. There are various clustering techniques that can be used for this purpose. In this work, we employ the hierarchical clustering \cite{7} as an algorithm of choice because it is well-known, and allows us to cluster into a set of groups the cardinality of which we do not know in advance. The clustering process is illustrated in Algorithm 1.

In the first stage of the clustering, each agent is regarded as belonging to a separate cluster, and the two clusters with the shortest 2-norm (Euclidean) normalised feature distance are then merged to form a new cluster. In the second stage, the merging of clusters continues as in the first stage, until either all the agents are assigned to a single cluster, or the $\delta^x_{G^i_T}$ of a potential group exceeds the feature threshold $fT$. At each stage of clustering, the expected behaviour $\delta^x_{G^i_T}$, of a potential group, is validated against the $bT$ threshold based on available evidence, to ensure the behaviour threshold is not exceeded.

Our clustering approach has some interesting characteristics. It imposes restriction on group membership for outlier agents. An outlier agent has features matching that of a particular group, but with a non-conforming behaviour to the group. Our model regards such agents as belonging to singleton groups pending evidence suggesting otherwise. In line with this, unknown agents start off in singleton groups even though their features may be matched to any of the existing groups, until there is sufficient evidence supporting their group membership.

2.4 Sampling and Evidence Aggregation

The DM model offers rich context from which an aggregation set may be derived. In the general case, an aggregation set is made of candidates randomly selected from different groups, from which evidence may be drawn in order to form an opinion. However, depending on the specific requirements of a task, richer contexts could be explored using the learned model of diversity. For instance, the cost and risk assessments of a potential transaction \cite{4}, may serve to inform the...
Algorithm 1 Hierarchical clustering algorithm for group formation using feature and behaviour criteria.

Require: A set of agents \( A \)
Require: A feature and behaviour based similarity thresholds, \( fT, bT \)
Require: A feature and behaviour similarity functions, \( \delta_{GF}, \delta_{GB} \)

1: allocate each agent in \( A \) as a single cluster
2: let \( C \) be the set of clusters
3: continue \( \Leftarrow \) true
4: while continue
5:    continue \( \Leftarrow \) false
6:    for all \( X, Y \in C \) do
7:        compute the between-cluster similarity \( \delta_{GF}(X,Y) \), such that \( \delta_{GB}(X,Y) < bT \)
8:    end for
9:    if \( fT < \delta_{GF}(X,Y) \) then
10:       \( Z \Leftarrow X \cup Y \), where \( \delta_{GF}(X,Y) \) is the minimum
11:       remove \( X \) and \( Y \) from \( C \)
12:       \( C \Leftarrow C \cup Z \)
13:       continue \( \Leftarrow \) true
14:    end if
15:  end while
16: return \( C \)

sampling process. Members in a group comprising of trustworthy agents may be favoured, for example, over agents in less trustworthy groups in a high-risk transaction. Also, in situations where the cost associated with sampling from specific groups of agents (e.g. groups of experts) exceeds a budget, groups of less knowledgeable agents may be considered, who in combination may provide a sufficiently similar service. We consider the random selection of one representative candidate from each of the groups to form an aggregation set. Provided the likelihood of agents in each of the groups behaving in a similar manner is relatively high, then evidence from the set may be considered a sufficient representation of the entire population. We do not suggest this to be the only approach for sampling, but only that it demonstrates one possible realisation of our model, which we have used in our evaluation. Other sophisticated sampling techniques may be explored to meet specific requirements.

We denote by \( S \) the aggregation set comprising of candidates sampled from groups in \( G \). \( G_{i, \text{last}} \) is the number of agents in \( G_i \). Further, we define \( G_{i}(l) \) as the index of the \( l \)th element in \( G_i \), for \( l = 1, \ldots, \text{last} \). A candidate \( y \in G_i \) is selected to be added to \( S \), by choosing a random integer \( z \in [1, \text{last}] \).

The DM model does not limit the chances of unknown agents being sampled by simply assigning them to an existing group having members with similar features. Every unknown agent, as already mentioned is regarded as belonging to its individual group, until there is sufficient evidence to classify it differently. This approach prevents a stereotypical treatment for such agents with regards to group membership, but gives each agent in this category a fair chance of being heard. There are benefits to this: in the first instance, a benevolent agent sharing similar features with a group of malicious agents will not be automatically labelled malicious, when there is no concrete evidence suggesting such. Also, a group of benevolent agents will not risk the abuse of its reputation by
malevolent agents who, for example, may be masquerading by presenting similar features [6].

An evaluator \( x \), combines reports from agents in the aggregation set \( S \subseteq A \), in order to arrive at an opinion about \( \rho \). Every agent \( s \), in the aggregation set, has its report weighted by the subjective trust value \( \tau \), assigned its group \( G_{i(s)} \) by \( x \). We use an a priori trust for unknown agents which is often set at 0.5 in literature [2]. The combined evidence is computed as:

\[
E^x_\rho = \frac{\sum_{s \in S} R_{y,\rho} \times \tau^{x}_{G_{i(s)}:\rho}}{\sum_{s \in S} \tau^{x}_{G_{i(s)}:\rho}}. \tag{9}
\]

3 Evaluation

In this section we describe experiments conducted to evaluate (in simulation) the performance of the Diversity Model. The aim of the experiments is to study the effect of group behaviour and deception on an aggregation result, and how these mechanisms may be exploited to limit the number of agents queried for evidence. We describe the methods adopted in the experiments, and present our results and discussion. The factors taken into account in the evaluation are: the predictive accuracy of the model to some ground truth, and the proportion of agents in the population sampled for evidence. We compare our approach to other approaches such as sampling the entire population of agents, randomly sampling a number of agents, and sampling based on the trustworthiness of the agents (in this instance we compare the performance of our model to the trust computation engine used in Beta Reputation System [5]).

3.1 Experimental Setup

Our experiments are based on a simulation testbed which models a logical network of agents as defined by their features. The environment consists of 100 sources and one evaluator. The evaluator relies on evidence obtained from the sources to evaluate a subject of interest. Our network is connected, with undirected edges from each node to its neighbours. The network fitness is based on the distance between features of agents, such that nodes that are highly similar gravitate towards each other. For simplicity, we assume that similarity is defined on the same feature dimension for all agents. Agents possess incomplete knowledge, and therefore, report with some amount of uncertainty. We simulate this phenomenon by drawing each agent’s report from a Gaussian distribution \( N(0, 1) \). However, agents closer to each other report in a similar manner. To simulate this we have each agent broadcast its report at each sampling phase to all its one-hop neighbours. Each node maintains a buffer of reports received from its neighbours. At each sampling phase, a node reports following \( N(0, 1) \). However, if there are corresponding reports in a node’s buffer for the same sampling phase,
a node alters its report to reflect a conformity to reports of its neighbours, with only a slight deviation. In this way, we define the underlying logical network we wish our evaluator to identify and exploit. The experimental parameters are listed in Table 1.

Table 1. Experimental Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># Information sources</td>
<td>100</td>
<td># evaluators</td>
<td>1</td>
</tr>
<tr>
<td>feature threshold ($fT$)</td>
<td>0.8</td>
<td>behaviour threshold ($bT$)</td>
<td>0.9</td>
</tr>
<tr>
<td>degree of nodes</td>
<td>8</td>
<td>report distribution</td>
<td>$N(0, 1)$</td>
</tr>
</tbody>
</table>

3.2 Experiments and Results

We consider different experimental conditions to analyse the effects of group behaviour and malicious agents on the aggregation result. We indicate the number of agents used by our model in each case to arrive at a decision. Each evaluation condition was initialised with random models of the information sources. 100 runs were conducted in 10 rounds for each case and the mean of the runs reported.

Effect of Group Behaviour In this experiment, we analyse the effect of the increasing rate of group behaviour in the population in the predictive accuracy of the evaluator agent. In real world scenario, subgroups may arise, for instance, as a result of agents having similar expertise, being constrained by organisational policies, or by a coordinated act of collusion by different agents. This may be regarded as a kind of deception, since agents exhibiting some of these traits may not be reporting their perception objectively. We linearly increase the percentage of conformity to group behaviour from 0% to 100%. In each case, agents that conform to group behaviour are selected randomly from the population. Figure 2 shows the effect of group behaviour on the aggregated result. The deviation from the ground truth is reported in each case. Although each agent reported with some uncertainty, given incomplete knowledge on a subject the evaluator took advantage of the large number of agents queried, and was able to make better predictions because the noise in the aggregated reports cancelled out, leaving only reliable reports. However, when the rate of group behaviour increased, as expected, the reports also became skewed in favour of opinions held by different subgroups, leading to lower accuracy in predictions. An evaluator in such circumstance may no longer benefit from sampling large number of agents, as each new report may only be a repetition of an already sampled opinion.

In Figure 2, the performance of the DM model is compared against other approaches. The metric we are specifically interested in is the performance of our Diversity Model compared to other approaches. We considered an approach based on sampling all agents in the population, which we refer to as the baseline. Also considered are models based on the trustworthiness of agents, and
the random selection of agents. The same number of agents as that sampled by the DM model was employed to select agents when using the trust model and the random selection respectively. The trust model involved sampling the most trustworthy agents in the population. Our first observation is that all the models begin very well when there were no group behaviour (at 0%). The predictions made were closer to the ground truth, with the baseline model slightly outperforming the other approaches. Also nearly as many agents as the baseline approach were queried by the DM model. This may be regarded as the worst case, where no compelling evidence could be established for the formation of informative groups. However, when evidence of group behaviour in the agent population emerged, our model was able to exploit this to reduce the number of agents queried, while still making better predictions. The performance of the trust model was worse off, undoubtedly caused by the uncertainty in the reports of the agents in each sampling phase. Specifically, when there were no expert or malicious agents in the system, the trust model was unable to learn any useful pattern in the reliability of agents, and therefore consistently made poor choices. The approach based on random selection of sources is an uninformed selection strategy, which leans much on chance. This, as observed is likely to perform poorly in environments where there are defined patterns of behaviour among agents. This observation is encouraging, as it demonstrates the efficacy of the DM model in guiding decision making rather than relying on chance.

![Fig. 2. Effect of group behaviour in aggregated result](image)

**Effect of Malicious Sources** Until now, we have discussed the scenario in which agents reported objectively based on incomplete knowledge. However, in real life settings, agents may not always behave benignly. There may be incentives for agents to lie, leading to distorted reports aimed at subverting the system. In this section we consider an attempt by malicious agents to systematically distort the aggregation result, by reporting a value different from their observation. Our goal was to determine robustness of our model with varying degrees of deception. As before, we compare the performance of the Diversity Model against the baseline approach, random selection, and trust filtering. The baseline, as in previous case, involved sampling all the agents in the population.
In the experiment, malicious agents report with a distribution that is different from normal agents. Deceptive reports were drawn from a Gaussian distribution $N(2, 0.01)$, with an attempt at distorting the aggregation result. However normal agents continued to report according to $N(0, 1)$. We gradually increased the percentage of deception in the system from 0 to 100, and observed the effect when group behaviour was kept constant at 20% and 80%, respectively. In each case, performance of the four approaches was considered, and the number of agents queried at each instance was also recorded. The number of agents queried by the diversity model in each case is captured in the result against the diversity sampling curve, as shown in Figure 3 and Figure 4.

An interesting observation, given this set of experiments is the notable improvement in the predictive accuracy of the trust model. Although still outperformed by the Diversity Model, the trust approach performed significantly better than the baseline and random approaches. The trust model, in this instance, was able to learn from its experience when the activity of malicious agents became obvious and stable in the system. The effect of deception is all the more highlighted and amplified in the system when considered in parallel with group behaviour. An analogy of this could be drawn to agents in a social network, where agents may be influenced based on the kind of social group they belong to, or who they listen to. This kind of phenomenon is referred to as rumour spreading in the literature [10]. As observed, the DM model could still make better predictions because it was able to adjust the weights for different types of sources.

4 Conclusion and Future Work

We have presented a framework for selecting sources of reputational evidence, in a way that guarantees reliable decisions from small groups of individuals. The approach presented in this work is oriented towards resource-constrained envi-
environments where querying of information sources is costly. However, the proposed approach could also be extended to other environments to facilitate selection of interaction partners, and to guard against deception, especially the more coordinated attempts of collusion. Where hidden networks defining group behaviour exist in the population, our model is able to exploit this in order to limit the number of information sources sampled while still remaining robust to deception. Where a naïve approach of evidence aggregation would perform poorly under these conditions, our model shows positive results that outperforms classical trust approach.

This work exploits features and perceived behaviour of agents in order to cluster them into groups. We intend to exploit richer information contexts such as domain ontology and provenance in future research in order to form better clusters. For simplicity, this work assumes that malicious agents behave in a consistent manner. We hope to incorporate more complex deception models in order to evaluate robustness of our model in other scenarios. Although we considered a very simple sampling mechanism for the selection of information sources, we intend to incorporate richer sampling techniques, aimed at satisfying different information needs of an application.

Acknowledgements. This research was supported by the Petroleum Technology Development Fund (PTDF) Nigeria, Overseas Scholarship Scheme (OSS).

References

A Subjectivity Alignment Approach for Effective Reputation Computation

Hui Fang, Jie Zhang, Murat Sensoy†, and Nadia Magnenat Thalmann

School of Computer Engineering, Nanyang Technological University, Singapore
†Department of Computing Science, University of Aberdeen, United Kingdom
{hfang1}@e.ntu.edu.sg

Abstract. Current reputation systems simply aggregate numerical ratings provided by buyers, but overlook the buyers’ subjectivity difference in evaluating the transactions with a seller. To address this problem, we propose a subjectivity alignment approach for reputation computation (SARC). It first requires the buyers to provide ratings and detailed reviews containing values of objective attributes of the transactions. After that, SARC applies Bayesian learning to model the correlations between each rating level and each objective attribute, and adopts a regression analysis model to learn the weights of the attributes, representing each buyer’s subjectivity. Ratings provided by one buyer can then be aligned (converted) for another buyer according to the two buyers’ subjectivity. Evaluation results indicate that SARC can more accurately and stably model sellers’ reputation than the BLADE and TRA VOS approaches. It is also not much affected by deception from dishonest buyers, and more robust to dynamic environments.

Categories: I.2.11 [Distributed Artificial Intelligence]: Intelligent agents; K.4.4 [Electronic Commerce]: Trust, Reputation

General Terms: Algorithms; Design

Keywords: subjectivity alignment, bayesian learning, reputation system, multi-agent system

1 Introduction

In open e-marketplaces, it is not possible for a buyer to have experience with every seller. On the other hand, dishonest sellers may advertise perfect deals but never deliver the promise. Therefore, there is a significant risk for buyers when selecting a seller among many alternatives. To address the issue, reputation systems [7] have been proposed, where buyers who previously bought products from a seller share their experience, normally in the form of a numerical rating reflecting the level of satisfaction about the transactions with the seller. These ratings are aggregated to represent the seller’s reputation. Other buyers can then rely on the reputation values of sellers to make decisions on which sellers to do business with.

A rating is a subjective evaluation of a seller by a buyer within the context of a specific transaction. Therefore, different ratings could be given for the same transaction by different buyers. Subjectivity difference may come from two sources. First, when the
buyer evaluates her satisfaction level with a transaction, she considers each attribute related to that transaction. Although the information about each attribute is objective, the evaluation (i.e., satisfactory level) of the attribute value may be subjective and change from user to user. This is referred to as intra-attribute subjectivity in this paper. For example, a product with the price of “USD1500” may be expensive for buyer a, while not so expensive for buyer b. Second, when the buyer assigns a satisfaction level to a transaction, she may consider some attributes of the transaction more heavily than others. This is referred to as extra-attribute subjectivity. For example, a buyer with better economic conditions may consider a product’s quality more heavily, while another buyer with worse economic conditions may concern more about the price of the product. The above two aspects together contribute to the subjectivity difference among buyers. Due to the subjectivity difference, it may not be effective if a buyer directly aggregates other buyers’ ratings to compute seller reputation. The computed reputation values may then mislead the buyer in selecting business partners.

To effectively address the subjectivity difference problem, we propose a subjectivity alignment approach for reputation computation (SARC). In our approach, each buyer is equipped with an intelligent (buying) agent. At the beginning of her interactions with the reputation system, a buyer a is required to provide her buying agent with both a single rating and a detailed review containing values of the objective attributes of transactions with sellers, such as price and delivery time, for each of a few transactions. Based on these rating-review pairs, the buying agent applies a proposed Bayesian learning approach to model the correlations between buyer a’s each rating level and the value of each objective attribute involved in the transactions. The learned correlation function, which represents buyer a’s intra-attribute subjectivity, will then be shared with the agents of other buyers. The agent of buyer a also applies a regression analysis model to learn the weight of each attribute for buyer a, representing her extra-attribute subjectivity. This information will not be shared with other buyers. After the learning phase, buyer a only needs to provide ratings for her transactions with sellers, not detailed reviews.

When another buyer b shares a new rating of her transaction with a seller, the agent of buyer a will first retrieve a rating level for each attribute of the transaction based on the shared rating and the intra-attribute subjectivity of buyer b shared by the agent of b. The rating levels of the attributes will then be aggregated according to buyer a’s extra-attribute subjectivity learned by the agent of a. In this way, the rating shared by buyer b is aligned to that can be used by buyer a for computing the reputation of the seller.

To evaluate the performance of our SARC approach, we simulate an e-commerce environment involving a number of buyers with different subjectivity in evaluating products and a set of sellers selling products with different attribute values. In addition, buyers’ subjectivity may change over time, buyers may also intentionally lie about their evaluation of products, and sellers may change the attribute values of their products. Experimental results confirm that our SARC approach provides sufficiently good performance in a general setting. It can more accurately and stably model sellers’ reputation than the representative competing approaches of BLADE [6] and TRAVOS [8]. Our approach is not dramatically affected by deceptive buyers because it treats dis-
honest buyers as the ones with different subjectivity. It is also more robust to dynamic environments.

2 Related Work

Quite a lot of filtering approaches have been proposed to address the problem of subjectivity difference among buyers and unfair ratings intentionally provided by dishonest buyers to mislead other buyers. For example, some of the approaches filter out the ratings of some buyers (advisors) whose past ratings differ significantly from the ratings of all advisors [9, 1], the ratings of a particular buyer [8, 10, 5], or the ratings of both [11]. These filtering approaches generally suffer from the risk of losing or discounting some important information. In contrast, our approach aligns/converts the ratings of the advisor to those that can be directly used by buyers according to the subjectivity of the buyers and the advisor learned by their agents.

Some other alignment approaches have also been proposed to align advisors’ advice about the trustworthiness of sellers. For example, Koster et al. [3] propose a trust alignment approach based on the general framework of Channel Theory. In this approach, each agent computes its own user’s trust evaluation patterns based on the interactions towards the same sellers (i.e., shared interactions). Then, the generalized patterns are used to align trust advice provided by advisors. The BLADE approach of Regan et al. [6] applies Bayesian learning to model sellers’ properties and the correlations between sellers’ properties and buyers’ ratings. Once a buyer receives a rating from an advisor, she can infer back the target seller’s properties, and then compute the rating of her own towards the seller on the basis of the inferred properties of the target seller. One shortcoming of these alignment approaches is that they ignore the intra-attribute subjectivity difference among buyers. Another shortcoming is that they require the buyer and the advisor to have shared interactions, which may not be the case in an e-commerce environment with a large population of sellers. In addition, these approaches generally offer limited flexibility for buyers to deal with the dynamic behavior of sellers and dynamic subjectivity of advisors. In contrast, our SARC approach aligns each rating provided by an advisor towards a transaction with a seller other than an aggregated trust value of the seller. In this way, it is not affected by sellers’ changing behavior. Our SARC approach updates the learned subjectivity of buyers (advisors) in certain interval of time to cope with the possible dynamic subjectivity of advisors. Our SARC approach does not rely on shared interactions. Instead, the agent of each buyer makes use of the ratings and detailed reviews provided by the buyer about her transactions with any sellers, to learn the buyer’s intra-attribute and extra-attribute subjectivity.

Collaborative filtering [4] and matrix factorization [2] have been proposed to address the subjectivity difference problem in the domain of recommender systems. However, recommender and reputation systems are different in the sense that reputation systems concern about sellers who may change behavior over time whereas recommender systems concentrate on static products. In addition, in reputation systems, a buyer may have several ratings towards one seller whereas a user has only one rating for one product in recommender systems.
3 The SARC Approach

In an e-marketplace, we denote the set of buyers by \( B = \{ b_1, b_2, b_3, \ldots \} \). The set of agents (called buying agents) equipped by corresponding buyers is denoted by \( A = \{ a_1, a_2, a_3, \ldots \} \), and the set of sellers are referred to as \( S = \{ s_1, s_2, s_3, \ldots \} \). The set of objective attributes for describing a transaction between a buyer and a seller is denoted as \( F = \{ f_1, f_2, \ldots, f_m \} \), where \( m \) represents the total number of objective attributes. Each rating provided by a buyer for a seller is from a set of predefined discrete rating levels \( L = \{ r_1, r_2, \ldots, r_n \} \), where \( n \) is the total number of different rating levels (i.e., the granularity of rating scale).

For a buyer \( b_i \in B \), the goal of her buying agent \( a_i \in A \) is to accurately compute the reputation value of a target seller \( s_j \in S \), according to \( b_i \)'s subjectivity. In order to achieve this goal, the buying agent \( a_i \) needs to consider the ratings of other buyers (advisors) that evaluate the satisfaction levels about their past transactions with seller \( s_j \). Due to the possible subjectivity difference between buyer \( b_i \) and the advisors, agent \( a_i \) also needs to align/convert ratings of each advisor (for example \( b_k \)) using our SARC approach.

More specifically, at the beginning of buyer \( b_i \)'s interactions with the system, agent \( a_i \) asks \( b_i \) to provide a rating for each of her transactions with a seller (which can be any seller in \( S \)). Buying agent \( a_i \) also asks \( b_i \) to provide detailed review information about each transaction containing the values of the set of objective attributes in \( F \). Based on the provided information (rating-review pairs), agent \( a_i \) models a set of correlation evaluation functions (CEFs) for buyer \( b_i \), capturing \( b_i \)'s intra-attribute subjectivity. Each correlation evaluation function is represented by a Bayesian conditional probability density function that models the correlation between each rating level and each objective attribute. Thus, for each buyer, the total number of the correlation evaluation functions is equal to \( m \times n \).

The learned CEFs of buyers will be shared with each other buyer’s agent. For a rating provided by the buyer (advisor) \( b_k \), agent \( a_i \) can then derive a rating for each attribute, based on the CEFs shared by \( b_k \)'s agent \( a_k \) and those of buyer \( b_i \)'s own. Note that what is derived for an attribute is in fact a set of probability values, each of which corresponds to a rating level in \( L \). The rating level with the highest probability will be chosen as the rating for the attribute.

Based on the provided rating-review pairs by \( b_i \), agent \( a_i \) also learns the extra-attribute subjectivity of buyer \( b_i \), which is represented by a set of weights for corresponding attributes in \( F \). The weight of an attribute is determined by two factors: 1) the probability value of the rating derived earlier; and 2) the importance of the attribute learned using a regression analysis model. These weights will not be shared with other buyers. Once the weights are learned, the aligned rating from that of advisor \( b_k \) can be computed as the weighted average of the derived ratings for the attributes.

In the next sections, we will describe in great details how SARC models CEFs based on rating-review pairs, derives a rating for each attribute, learns the weights for attributes, and computes a (aligned) rating by aggregating the derived ratings for attributes, organized as intra-attribute subjectivity alignment and extra-attribute subjectivity alignment.
3.1 Intra-attribute Subjectivity Alignment

Given a set of rating-review pairs provided by buyer $b_i$, each of which is for a transaction between $b_i$ and a seller, the rating in a pair indicates $b_i$’s satisfaction level about the corresponding transaction, and the review in the pair is a set of values for the attributes $F$ of the transaction. Buyer $b_i$’s agent $a_i$ learns the correlation evaluation functions (CEFs) of $b_i$, each of which is represented by a Bayesian conditional probability density function. Each CEF is the correlation between a rating level and the values of an attribute. More specifically, let us learn CEF $b_i^{u,v}$, the correlation function between attribute $f_u$ and rating level $r_v$ for buyer $b_i$, where $1 \leq u \leq m$ and $1 \leq v \leq n$. Buying agent $a_i$ first learns $p_{b_i}(r_v)$ (the probability that buyer $b_i$ provides a rating $r_v$), $p_{b_i}(f_u)$ (the probability distribution of the values for attribute $f_u$), and $p_{b_i}(r_v | f_u)$ (the conditional probability of rating level $r_v$ given the distribution of the values for attribute $f_u$).

By applying the Bayes’ Rule, agent $a_i$ can derive CEF $b_i^{u,v}$ as the conditional probability distribution of the values for attribute $f_u$ given rating level $r_v$ as follows:

$$CEF_{u,v}^{b_i} = p_{b_i}^{f_u}(r_v | f_u) = \frac{p_{b_i}(r_v \mid f_u) \times p_{b_i}(f_u)}{p_{b_i}(r_v)}$$

In our approach, the agents of buyers share the learned CEFs for their buyers with the agents of other buyers. Suppose that the agent $a_k$ of a buyer $b_k$ shares the learned CEF $b_k^{u,v}$ for $b_k$ with the agent $a_i$ of buyer $b_i$. For a rating $r_{b_k}$ shared by buyer $b_k$, agent $a_i$ can then derive a rating level for each attribute in $F$. We use a Naïve Bayesian Network model to learn the mapping/alignment from $r_{b_k}$ of buyer $b_k$ to the ratings of $b_i$ for the attributes, as illustrated in Figure 1. Although in this model we assume that the attributes are independent given the ratings of buyers, in the next section, we will learn the relative weights of the attributes to capture the dependency among the attributes.

Let us take any $f_u \in F$ as an example attribute to show how agent $a_i$ derives a rating for attribute $f_u$. To do so, agent $a_i$ first estimates the conditional probability of a rating level in $L$ for attribute $f_u$, given rating $r_{b_k}$ provided by buyer $b_k$. Take any rating level $r_{b_k}$ as an example, agent $a_i$ computes $p_{b_i}(r_{b_k} | f_u)$, the conditional probability that buyer $b_i$ will assign the rating level $r_{b_k}$ to attribute $f_u$ given the rating $r_{b_k}$ of
buyer $b_k$, as follows:

$$p_{b_i}^h(r_v,f_u | r_{b_k}^h) = \frac{p_{b_i}^h(r_v | f_u, r_{b_k}^h) \times p_{b_i}^h(f_u | r_{b_k}^h)}{p_{b_i}^h(f_u | r_{b_k})}$$

where $p_{b_i}^h(f_u | r_{b_k}^h)$ is learned by agent $a_k$ of buyer $b_k$ using Equation 1 and shared by agent $a_k$ to agent $a_i$. $p_{b_i}^h(f_u | r_v)$ is learned by $a_i$ itself using Equation 1, and $p_{b_i}^h(r_v | f_u)$ is obtained by agent $a_i$ from the rating-review pairs provided by its buyer $b_i$. In Equation 2, $p_{b_i}^h(r_v | f_u, r_{b_k}^h)$ is equivalent to $p_{b_i}^h(r_v | f_u)$ and $p_{b_i}^h(f_u | r_v, r_{b_k}^h)$ is equivalent to $p_{b_i}^h(f_u | r_v)$ because buyer $b_i$ provides ratings to corresponding attributes regardless of buyer $b_k$’s ratings. In other words, buyers evaluate transactions independently.

For attribute $f_u$, agent $a_i$ learns the conditional probability of each rating level $r_v \in \mathcal{L}$ according to Equation 2. The aligned rating of attribute $f_u$ for buyer $b_i$ on the basis of buyer $b_k$’s rating is then determined as the rating level with the highest probability value, as follows:

$$r_{u,k}^{b_i} = \arg\max_{r_v \in \mathcal{L}} (p_{b_i}^h(r_v, f_u | r_{b_k}^h))$$

The aligned ratings for other attributes in $\mathcal{F}$ can also be determined in the same way according to Equations 2 and 3.

### 3.2 Extra-attribute Subjectivity Alignment

After the ratings of the attributes are obtained, agent $a_i$ of buyer $b_i$ then aggregates the ratings to represent an aligned rating of the rating $r_{b_k}^h$ shared by buyer $b_k$. To do this, $a_i$ needs to first determine a weight for each attribute in $\mathcal{F}$ as buyer $b_i$ may concern more about one attribute over another.

The weight of an attribute $f_u$ is determined by two factors. One factor is the confidence $C_u$ about the rating $r_{u,k}^{b_i}$ derived for the attribute $f_u$ using Equations 2 and 3. The confidence can be represented as the conditional probability value of the derived rating, $p_{b_i}^h(r_{u,k}^{b_i}|r_{b_k}^h)$ estimated using Equation 2. A larger probability value means that it is more probable that the derived rating for attribute $f_u$ should be $r_{u,k}^{b_i}$ according to buyer $b_k$’s rating and the subjectivity of buyers $b_i$ and $b_k$. In another word, the larger the probability is, the more reliable the derived rating $r_{u,k}^{b_i}$ is. Thus, we have:

$$C_u = p_{b_i}^h(r_{u,k}^{b_i}|r_{b_k}^h)$$

Another factor to determine the weight for attribute $f_u$ is the importance $I_u$ of $f_u$ in buyer $b_i$’s view. The importance $I_u$ can be modeled as the coefficient of attribute $f_u$ by a regression analysis model, based on the rating-review pairs provided by $b_i$. More specifically, given the rating-review pairs, we compute the coefficients for attributes by minimizing the aggregated difference between the true ratings in the rating-review pairs of $b_i$ and the ratings, each of which is predicted for a review by the following equation:

$$r_0^{b_i} = I_0 + \sum_{u=1}^{m} I_u \times V_{f_u} + \varepsilon$$

(5)
where \( r^{b_i}_{0} \) is the predicted rating for a review, \( V_{fu} \) is the value of \( f_u \) in the review, \( I_0 \) is a constant, and \( \varepsilon \) is residual. So, the coefficients \( I = [I_0, I_1, \ldots, I_m] \) can be computed by:

\[
I' = (X'X)^{-1}X'Y
\]  

(6)

where if there are \( c \) rating-review pairs for buyer \( b_i \) in total,

\[
X = \begin{bmatrix}
1 & f_{11} & \cdots & f_{m1} \\
1 & f_{12} & \cdots & f_{m2} \\
\vdots & \vdots & \ddots & \vdots \\
1 & f_{1c} & \cdots & f_{mc} \\
\end{bmatrix}, \quad Y = \begin{bmatrix}
r_1 \\
r_2 \\
\vdots \\
r_c \\
\end{bmatrix}
\]

After the weight (confidence and importance) of each attribute is determined, the aligned rating \( r^{b_i}_{k} \) can be computed as the weighted average of the ratings for attributes derived using Equations 2 and 3, as follows:

\[
r^{b_i}_{k} = \frac{\sum_{u=1}^{m} r^{b_i}_{u,k} \times C_u \times I_u}{\sum_{u=1}^{m} C_u \times I_u}
\]  

(7)

After aligning all ratings shared by all buyers (advisors), the reputation value of seller \( s_j \) in the view of \( b_i \) can be computed as, for example, the average of the aligned ratings.

4 Experimentation

In this section, we carry out experiments to evaluate the performance of our SARC approach and compare it with some representative competing approaches. We simulate an e-commerce environment involving 50 sellers and 200 buyers. In our simulations, sellers may provide different products. Their products are all different PC configurations with five objective attributes, namely, \( \text{Price} \), \( \text{Speed of CPU} \), \( \text{Processor Type} \), \( \text{Graphics Card Type} \), and \( \text{Hard Drive Size} \) with ranges presented in Table 1. For each seller, the values of the five attributes of her products are randomly chosen within the ranges.

Buyers may have different subjectivity in evaluating their transactions with (the products of) sellers. We simulate both buyers’ intra-attribute subjectivity and extra-attribute subjectivity. To be specific, we assume that a buyer’s rating for a transaction with a seller is derived as follows. First, the buyer evaluates each objective attribute according to a specific intrinsic (taste) function. In our experiments, buyers’ intra-attribute subjectivity is simulated as an approximate \textit{Gaussian Distribution}. That is, for each attribute, the probability of each rating level given by a buyer is in the form of a normal distribution. Second, the buyer places random weights (in the domain of \([0,1]\)) on different attributes, and computes the weighted average of her evaluations on attributes as a single rating for the transaction. Since buyers can only give ratings under the predefined rating scale in reality, the simulated rating is chosen from the predefined rating scale that is the closest to the weighted average.

In the experiments, we also implement a baseline approach without subjectivity alignment, which computes seller reputation by directly averaging the ratings collected.
Table 1. Product Attributes and Value Ranges

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Type</th>
<th>Ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>Double</td>
<td>$100-$10,000</td>
</tr>
<tr>
<td>Speed of CPU</td>
<td>Double</td>
<td>1-10 GHZ</td>
</tr>
<tr>
<td>Processor Type</td>
<td>Char</td>
<td>5 types</td>
</tr>
<tr>
<td>Graphics Card Type</td>
<td>Char</td>
<td>2 types</td>
</tr>
<tr>
<td>Hard Drive Size</td>
<td>Integer</td>
<td>40-1000GB</td>
</tr>
</tbody>
</table>

from other buyers. We choose to implement TRAVOS [8], a representative filtering approach (see the Related Work section for details). BLADE [6] is chosen instead of the approach of Koster et al. [3] because they are very similar and the approach of Koster et al. is complicated to implement.

We compare the performance of these approaches with our approach in reputation computation. The performance of an approach is measured as the mean absolute error (MAE) between seller reputation computed for each buyer using the approach, and that using the ratings according to each buyer’s own subjectivity (representing the ground truth about seller reputation with respect to the buyer).

To simulate real-world e-commerce environments, we set several important parameters for our simulations, including information availability, dynamic behavior of sellers, dynamic subjectivity of buyers, ratio of liars (dishonest buyers), and granularity of rating scale.

Information availability refers to the amount of available information required by different approaches for subjectivity alignment. Two types of information are needed by our approach. One is the detailed reviews describing the objective attributes of transactions between buyers and sellers. This information is used by our approach to model the correlation evaluation functions (CEFs) and the importance of the attributes for buyers. We vary the number of detailed reviews ($N_r$) to see how the performance of our approach is affected by this parameter. Another type of information contributing to our approach is the number of objective attributes. In reality, some attributes (e.g. appearance) may not be objective. The total number of objective attributes in our simulations may thus be less than 5. We vary the ratio of objective attributes ($R_{obj}$) to be 0%, 20%, 40%, 60%, 80%, and 100%, to see how much the performance of our approach will be affected. One type of information required by the BLADE approach is shared interactions where buyers and advisors have interacted with some same sellers. We vary the ratio of shared interactions ($R_i$) to see how this parameter affects the performance of BLADE.

We also set the parameter $P_{seller}$ to capture the dynamic behavior of sellers. In real-world e-commerce environments, sellers may change their behavior over time. For example, they may provide products of high quality at first, but those of low quality after earning enough reputation. In our experiments, dynamic behavior of sellers is simulated by changing the quality of their products (i.e. the values of a subset of the objective attributes in Table 1).
Buyers may also adjust their subjectivity over time. Dynamic subjectivity of buyers ($P_{buyer}$) is captured in their rating procedure by adjusting intra-attribute subjectivity, or extra-attribute subjectivity, or both.

Ratio of liars ($R_{liar}$) is adopted to reflect the deception problem in real e-marketplaces where some buyers may lie about their experience with sellers. Following the work of [10, 9, 8], we also simulate the complementary lying behavior where if a true rating to a seller is $r$ in the scale of $[0, 1]$, the liar will modify the rating as $1 - r$.

Granularity of rating scale ($G_{scale}$) refers to the number of rating levels. It may be different for different reputation systems. In our experiments, we will study the effect of the granularity of rating scale by varying $G_{scale}$ from 2 to 10.

We vary the values of the above parameters to simulate basic, deceptive and dynamic environments, respectively.

### Basic Environment
We first simulate a basic environment without any variation of the parameters (i.e., $R_{liar} = 0$, $P_{seller} = 0$, $P_{buyer} = 0$), and compare the performance of our approach and that of the three competing approaches, including the baseline approach, TRAVOS and BLADE. We compute their mean absolute error (MAE) values for computing the reputation of sellers in different epoches. In each epoch, each buyer
interacts with one seller in the marketplace. From the results shown in Figure 2(a), we can see that our approach performs consistently the best no matter whether buyers have more or less experience with sellers. Because both TRAVOS and BLADE require shared interactions, their performance is limited. Both TRAVOS and BLADE perform slightly better than the baseline approach. The performance difference between the different approaches is reduced when buyers have more experience with sellers in the marketplace.

Based on the basic environment, we then vary some parameters to examine their effects. We first examine how the ratio of objective attributes $R_{obj}$ affects our SARC approach. We vary $R_{obj}$ from 0% to 100% for our SARC approach, while keep $R_{obj}$ to be 100% for BLADE. As shown in Figure 2(b), SARC performs slightly worse than BLADE when there are no objective attributes. However, it performs better than BLADE when there are more than 20% of objective attributes. The performance of SARC consistently increases as the ratio of objective attributes increases. But, the increment becomes smaller when $R_{obj} \geq 20\%$.

The larger the granularity of the rating scale ($G_{scale}$) is, the easier to learn buyers’ subjectivity because buyers’ subjectivity can be better captured by the larger grana-
larity of the rating scale. This trend is verified by our experiment. In Figure 2(c), we plot the MAE results of the four approaches when varying $G_{scale}$ from 2 to 10. The figure shows that the performance of SARC is significantly greater than the baseline approach, TRAVOS and BLADE. On average, the performance of SARC improves as $G_{scale}$ increases.

We also vary the number of detailed reviews ($N_r$) provided by buyers from 1 to 30. We try to figure out a reasonable $N_r$ for SARC. As shown in Figure 2(d), when $N_r$ increases from 1 to 5, the performance of SARC increases significantly. While $N_r$ is larger than 5, as the increase of $N_r$, the performance of SARC also increases, but in a much smaller degree. This is simply because SARC requires only a few detailed reviews to learn buyers’ subjectivity well. After that, any additional information leads to only small improvement. Thus, we can choose 6 as the acceptable minimum $N_r$. Besides, SARC performs better than the baseline approach and BLADE in all the cases for $N_r$.

BLADE requires shared interactions in order to learn buyers’ subjectivity. However, in real e-marketplaces, shared interactions are generally very sparse. In this experiment, we fix the number of past interactions for each buyer, but vary the ratio of shared interactions ($R_i$) from 0% to 100%. For each ratio value, MAE is computed as the average of five repeated runs. Figure 3(a) indicates that BLADE performs significantly worse than SARC when $R_i$ is in the range from 0% to 30%. The performance of BLADE increases with the increase of $R_i$.

**Deceptive Environment** In this experiment, we examine the effect of deception (buyers lying about their past experience) on different approaches. We vary the ratio of liars ($R_{liar}$) from 0% to 100%, and plot the MAE results of different approaches in Figure 3(b). We can see that the performance of TRAVOS does not decrease much as $R_{liar}$ increases. Our SARC performs much better than the other three models for any $R_{liar}$. It is not dramatically affected by lying buyers because SARC learns a buyer’s subjectivity from the buyer’s own past experience and treats lying buyers as buyers with different subjectivity. When $R_{liar}$ is larger than 0.5, BLADE performs worse than TRAVOS, but consistently better than the baseline approach. Note that in the environment where most buyers are liars, the performances of other models are not so bad. This is mainly because buyers have different subjectivity in our simulations. The effect of buyers’ lying behavior may be reduced by the subjectivity difference among buyers, and vice versa.

**Dynamic Environment** In this experiment, we simulate the environment where sellers may change the quality of their provided products in their transactions with buyers. We define a predefined parameter, $P_{seller}$, to represent the probability that each seller may vary the values of the five attributes of her provided products. We assume that sellers only change their behavior once in the marketplace. Once their behavior is changed, they will keep the behavior. $P_{seller}$ is ranged from 0 to 1 and increased by 0.05 in our experiment. The MAE results for SARC and other three approaches are plotted in Figure 3(c), which demonstrates that the performance of SARC is not sensitive to the dynamic behavior of sellers, and it performs almost consistently in all cases, while the performance of Baseline, TRAVOS and BLADE gets worse as the increase of $P_{seller}$.
The main reason is that SARC models the rating behavior (subjectivity) of each buyer from the buyer's own experience, which is independent of sellers’ behavior change. For TRAVOS and BLADE, they rely on past shared interactions between the buyer and advisors, and these shared interactions may not be suitable source information used for aligning the buyer’s subjectivity due to the possible behavior change of sellers in the shared interactions. For example, for a buyer and an advisor with the same subjectivity, if they interact with a seller in different time periods where the seller has changed behavior, TRAVOS may incorrectly treat the advisor as a liar and BLADE may incorrectly conclude that the buyer and the advisor have different subjectivity.

In a marketplace, buyers may also change or adjust their subjectivity after several interactions with sellers. In this experiment, we assume that buyers will change their subjectivity with a certain predefined probability, \( P_{\text{buyer}} \). Same as the previous experiment, buyers only change their subjectivity once in the marketplace and then keep their changed subjectivity in the following interactions with sellers. Figure 3(d) shows that the performance of SARC is not affected by buyers’ dynamic subjectivity. In SARC, buying agents can update the learned subjectivity of buyers by acquiring their buyers’ own recent experience, which provides flexibility to deal with their buyers’ dynamic subjectivity. The performance of BLADE becomes almost equivalent to that of Baseline as \( P_{\text{buyer}} \) increases, and is consistently lower than SARC. In BLADE, once a buyer’s subjectivity is changed, her buying agent cannot align ratings from advisors effectively because new shared interactions between the buyer and advisors are needed. TRAVOS performs worse than Baseline as \( P_{\text{buyer}} \) increases because the learned results of advisors become misleading after they change subjectivity.

5 Conclusion and Future Work

In this paper, we proposed a subjectivity alignment approach for reputation computation, SARC, to address the subjectivity difference problem. In SARC, buyers’ subjectivity is learned based on the ratings and detailed reviews they provide about the objective attributes of their transactions with sellers. More specifically, SARC separately learns the intra-attribute subjectivity and extra-attribute subjectivity of buyers. Buyers’ intra-attribute subjectivity is modeled using Bayesian learning. Their extra-attribute subjectivity is learned using a regression analysis model. We also conducted various experiments to compare the performance of our approach with that of other three competing models, including the baseline approach, TRAVOS and BLADE. Experimental results demonstrate that: 1) SARC performs better than the other three approaches, and can more accurately and stably model sellers’ reputation; 2) SARC is capable of coping with environments with deception and dynamic buyer and seller behavior; 3) the requirement of detailed reviews and objective attributes is not very restrictive.

For future work, we will conduct more experiments on real data obtained from, for example, eBay (www.ebay.com) to further validate the effectiveness of our approach.
6 Acknowledgement

This work has been made possible thank to the Institute for Media Innovation at Nanyang Technological University who has given a scholarship to the first author.

References

A Model for Augmenting Trust Management using Argumentation

Sharmin Jalal1, Karl N. Levitt1, Jeff Rowe1, Elizabeth Sklar2,3, and Simon Parsons2,3

1 Dept of Computer Science, University of California at Davis, CA, USA
sjalal@ucdavis.edu, {levitt, rowe}@cs.ucdavis.edu
2 Dept of Computer & Information Science, Brooklyn College, City University of New York,
2900 Bedford Avenue, Brooklyn, NY 11210 USA
{sklar, parsons}@sci.brooklyn.cuny.edu
3 Dept. of Computer Science, Graduate Center, City University of New York
365 Fifth Avenue, New York, NY 10016, USA

Abstract. Previous work has used theories of evidence to incorporate belief into trust and reputation systems. Some important questions that remain, however, are how agents might recover reputation lost in disputed transactions, and how new agents with little or no past transaction history might enter the trust network. We attempt to address these issues by extending previous work using the Dempster-Shafer theory of evidence to include formal argumentation. Reasons for past bad transaction assignment can be taken into account in new transactions and discounted by importance. New agents can participate in trust networks by forwarding evidence as arguments in a distributed reputation system. We present our preliminary model on incorporation of argumentation frameworks into trust management systems to support more complex reasoning mechanisms.

Keywords: Dempster-Shafer theory, argumentation, trust.

1 Introduction

Trust and reputation systems have gained widespread use and are increasingly important in distributed online applications. Online financial transactions, social networking and mobile ad hoc networks are some typical examples where trust is used to gauge the potential for successful exchange. A variety of research has modeled methods for building reputation and combining reputation into the notion of trust [9, 12, 14–17]. Defining and measuring the quality of trust, finding out domain suitable incentives to encourage participation, dealing with false identities that aim to deceive others by contaminating trust and decreasing trust transitivity are some potential research areas.

Most reputation systems assume the existence of a pre-built trust network, where some initial trust values are already in place. Very few provide a basis for deriving the initial value when past transaction records are unavailable. Another interesting area deserving attention is the recovery of reputation after a small number of anomalous bad transactions. How might a trust query distinguish between transactions with some strong reason why the outcome was bad and transactions whose outcomes were judged
bad for minor reasons? In these cases, the target agent deemed bad has no way to defend itself. Again, most reputation systems consider only “witnesses” (who participate in the transaction by giving reference) or “target agents” (about whom the query pertains). But there could be other agents that are neither targets nor reference providers, yet have some relevant information that could be very helpful in making decisions about trust. Currently, these agents and their information are mostly neglected in decision making.

We assert that argumentation is a mechanism which gathers both complete and incomplete information from different sources and reaches a conclusion through logical reasoning. Consider the situation when a seller in a financial transaction is tagged as untrustworthy and wishes to defend himself. Argumentation allows us to logically infer the reason behind a supposed bad transaction from the propositions exchanged between the buyer and the seller agents involved.

Here, we describe our preliminary work in using argumentation to address the above areas in reputation and trust management. We extend the work of Yu and Singh [22], which proposes a distributed reputation management system using the Dempster-Shafer theory of evidence. In the following sections, we describe the background work, followed with the extensions in our model.

2 Background

In [22], the authors proposed a reputation management system which employs the Dempster-Shafer theory of evidence as the underlying computational framework. Their model applies the Dempster-Shafer belief function and Dempster’s rule of combination to compute local and total belief of agents. Our model extends Yu and Singh’s model [22] and resolves the scenario when an agent wants to defend himself to retrieve his past good reputation. In addition, we propose a mechanism to aggregate discrete but relevant information from trusted agents and use this in measuring belief in a specific agent. We also discuss rewards and penalties to control the flow of authentic information between agents.

In Section 2.1, we give the basic notions of the Dempster-Shafer Theory of evidence, which is the foundation for Yu and Singh’s work and for our extensions. Then we describe how Dempster-Shafer theory was issued by Yu and Singh. In Sections 2.2 and 2.3, we elaborate very briefly on Yu and Singh’s way of computing “local trust” from past transactions and “total trust” by combining the local trust values of neighbors. Section 3 then introduces argumentation and describes how argumentation can be used to extend the kind of reasoning possible in Yu and Singh’s work. Finally, Section 4 summarizes and outlines future work.

2.1 Dempster-Shafer Theory

The seminal work on the Dempster-Shafer (DS) theory of evidence is Shafer’s work “A Mathematical Theory of Evidence” [20] which is an extension of Dempster’s work “Upper and Lower Probabilities induced by a Multivalued Mapping” [4]. We can say DS theory is a generalization of traditional probability theory, except that in DS theory, probabilities are assigned to sets of hypotheses instead of a single hypothesis. This
property makes DS theory more expressive than simple probability theory. In DS theory, there is no relationship between believing in a hypothesis and disbelieving it. Say agent A’s belief in some hypothesis is 0.8. According to DS theory, it is not necessary to assign the remaining 0.2 to be disbelief in the hypothesis, but rather it could be assigned to the set of all the possible hypotheses, indicating a lack of knowledge about them. As evidence is accumulated, the uncertainty narrows down to a subset of the entire hypothesis set [11]. Say we have two hypotheses $T$ and $\neg T$, then $Bel(T)$ represents belief in hypothesis $T$, $Bel(\neg T)$ represents belief in hypothesis $\neg T$, which is disbelief in $T$, and $Bel([T, \neg T])$ represents belief in the hypothesis $T$ or $\neg T$, which represents a lack of belief in $T$ or $\neg T$, or, alternatively, uncertainty about which of $T$ and $\neg T$ is true.

Another feature of DS theory is that it does not require a priori knowledge, which makes it appealing in cases with no previous data.

Below we introduce the terminology upon which we base our work.

**Definition 1 (Frame of Discernment).** The Frame of Discernment $\Theta$ is the set of exhaustive and mutually exclusive hypotheses under consideration.

While DS theory allows for arbitrary frames of discernment, in this paper we will typically be concerned with frames of discernment that contain just a proposition and its negation $\{T, \neg T\}$.

**Definition 2 (Basic Probability Assignment).** The Basic Probability Assignment (BPA) is a function mapping the power set of the frame of discernment to the interval between 0 and 1. The BPA of the null set is 0 and the summation of BPA’s of all the subsets of the power set is 1.

We can write the constraints on the basic probability assignment as follows,

$$m : 2^\Theta \rightarrow [0, 1]$$

where $\Theta = \{T, \neg T\}$ is the frame of discernment. We will write $\mathcal{L}$ for $2^\Theta$, and so we have:

$$m(\emptyset) = 0$$

and

$$\sum_{A \in \mathcal{L}} m(A) = 1$$

Thus:

$$m([T]) + m([\neg T]) + m([T, \neg T]) = 1$$

$m(A)$ is also called the basic probability number and is the measure of the belief that is committed exactly to $A$ and does not include any belief committed to any subsets of $A$.

**Definition 3 (Belief Function).** For a subset $A$, the Belief Function $Bel(A)$ sums the basic probability number, or total belief, of all the nonempty subsets of $A$ which are also called the Focal Elements of $Bel(A)$. The Belief Function of $A$ is defined as:

$$Bel(A) = \sum_{B \subseteq A} m(B)$$
Thus,
\[ \text{Bel}(T, \neg T) = m(T) + m(\neg T) + m(T, \neg T) = 1 \]

For individual members, Bel and m are the same. Therefore, \( \text{Bel}(T) = m(T) \) and \( \text{Bel}(\neg T) = m(\neg T) \).

### 2.2 Local Belief from Statistical Data

[22] gives two ways to evaluate the trustworthiness of a given agent, called the target. The first is used when some other agent has sufficient previous experience with the target agent. In this case, the agent will build its local belief towards the target from the historical data. The process is as follows: after each transaction, the agent collects its user’s rating about the transaction and saves the latest, say, \( H \) of them. Suppose agent \( A \) has had several past transactions with agent \( V \), and he wants to evaluate the trust he assigns to \( V \). Thresholds \( \Omega \) and \( \omega \) are defined as the upper and lower trust limits of agent \( A \) respectively. Function \( f(\rho) \) returns the probability of a given value \( \rho \), where \( \rho \in \{0.0, 0.1, 0.2, \ldots, 1.0\} \) represents the quality of the services reflected in past transaction ratings for \( V \). A’s local belief towards \( V \), according to [22] is the following:

\[
\begin{align*}
\text{Bel}(T) &= m(T) = \sum_{\rho=\Omega}^{\rho=\omega} f(\rho) \\
\text{Bel}(\neg T) &= m(\neg T) = \sum_{\rho=0}^{\rho=\omega} f(\rho) \\
\text{Bel}(T, \neg T) &= m(T, \neg T) = \sum_{\rho=\Omega}^{\rho=\omega} f(\rho)
\end{align*}
\]

### 2.3 Combining Beliefs of the Witnesses

The second approach to evaluating trustworthiness in [22] is collecting local belief from the witnesses. Suppose that \( A \) does not have many past transactions with the target \( V \). In this model, \( A \) will ask for references from its trusted neighbors. If they have had enough transactions with \( V \), they will already have computed their local trust and can pass that value to \( A \). If, however, they also lack sufficient transaction history, they will pass a reference to another of their trusted agents in turn. The referenced agent then supplies its local trust about \( V \) or passes along yet another reference. The authors define a depthLimit as the maximum length of the referral chain. Let \( \alpha \) be the focal element of belief function \( \text{Bel} \) over \( L \). \( \text{Bel1} \) and \( \text{Bel2} \) are two belief functions over \( L \) based on different evidence. \( m1 \) and \( m2 \) are the BPA’s of \( \text{Bel1} \) and \( \text{Bel2} \), respectively. According to Dempster’s rule of combination, \( m = m1(\alpha) \oplus m2(\alpha) \) will be the new combined BPA over \( \alpha \), which is the sum of the form \( m1(X)m2(Y) \), where \( X \) and \( Y \) range over all subsets whose intersection is \( \alpha \). Therefore,

\[ m(\emptyset) = 0 \]
Here, \( \{X_1, X_2, \ldots X_n \} \) are the focal elements of \( m_1 \), and \( \{Y_1, Y_2 \ldots Y_m \} \) are the focal elements of \( m_2 \). And,

\[
\sum_{X_i \cap Y_j = \emptyset} m_1(X_i) m_2(Y_j) < 1
\]

is also called conflict. This indicates the conflict between two distinct bodies of evidence.

In the model, \( \tau \) and \( \pi \) are defined as the functions that return the local belief and total beliefs of an agent, respectively. Therefore, in the presence of the witnesses \( A = \{w_1, w_2, \ldots w_n \} \), agent \( A \) will update its total belief over \( V \), considering all of the local beliefs from its witnesses.

\[
\pi_A = \tau_{w_1} \oplus \tau_{w_2} \oplus \ldots \oplus \tau_{w_n}
\]

As threshold for trustworthiness is then defined. Agent \( A \) will trust agent \( V \) if,

I. \( \tau_A(T_V) - \tau_A(\neg T_V) \geq \text{trust threshold} \), in the case when agent \( A \) constructs its local belief from its own historical data.

II. \( \pi_A(T_V) - \pi_A(\neg T_V) \geq \text{trust threshold} \), when agent \( A \) constructs its total belief, combining the local belief of its witnesses.

Having described the approach suggested by Yu and Singh, we will go on to describe how argumentation can be used to extend the model.

3 Argumentation to Compute Trust

In the following sections, we will first describe the basic ideas of argumentation frameworks and the acceptability semantics. In later sections, we will describe our model in different scenarios.

3.1 Argumentation Background

In this subsection, we briefly describe some key elements of argumentation. We follow Dung’s notions of argumentation [5], where an argumentation framework is an abstract entity whose role is determined by its relation to other arguments.

**Definition 4 (Argumentation Framework).** An argumentation framework is a pair:

\[
AF = (AR, R)
\]

where \( AR \) is the set of arguments and \( R \) is the binary attack relation between arguments. That is, \( R \subseteq AR \times AR \).
For two arguments $A$ and $B$, we say $A$ attacks $B$ if $(A, B) \in R$.

To illustrate further the notion of argumentation, we are considering a particular argumentation system stated in [2] that handles inconsistency in the knowledge base. According to [2], arguments are built from a propositional knowledge base $\Sigma$ that could be inconsistent. $\vdash$ stands for classical inference and $\equiv$ stands for logical equivalence.

**Definition 5 (Argument).** [2] An argument is a pair $(H, h)$, where $H \subseteq \Sigma$ such that $H \vdash h$

$H$ is assumed to be consistent and minimal (for set inclusion). $H$ is called the support, and $h$ is the conclusion of the argument.

To illustrate the attack relation a little more, [7] defined two relations, Rebut and Undercut, which are as follows:

**Definition 6 (Rebut).** Let $(H_1, h_1)$ and $(H_2, h_2)$ be two arguments. $(H_1, h_1)$ rebuts $(H_2, h_2)$ iff $h_1 \equiv \neg h_2$.

**Definition 7 (Undercut).** Let $(H_1, h_1)$ and $(H_2, h_2)$ be two arguments. $(H_1, h_1)$ undercuts $(H_2, h_2)$ iff $\exists h \in H_2$ such that $h \equiv \neg h_1$.

Though the definition of attack in [2] includes the notion of rebut, we do not use rebut here because it has been shown to have some unfortunate consequences for argumentation systems using propositional logic [1].

**Definition 8 (Conflict Free).** We say, a set $S$ is conflict-free if $\forall A \in S$, $\not\exists B \in S$ such that $(B, A) \in R$.

**Definition 9 (Acceptable).** An argument $A$ is acceptable with respect to a set $S$ iff $\forall B \in AR$, if $(B, A) \in R$, then $\exists C \in S$ such that $(C, B) \in R$.

That is, an argument is acceptable to a rational agent, iff he can defend that argument from his own knowledge base.

**Definition 10 (Admissable).** Consider $S$ as a conflict-free set of arguments in the framework $(AR, Attacks)$. $S$ is admissible iff each argument in that set is acceptable with respect to set $S$.

**Definition 11 (Preferred Extension).** A preferred extension is the maximal (with respect to set inclusion) admissible set of the argumentation framework $AF$.

**Example 1.** Let $AF = \langle\{A, B, C\}, \{(B, A)(C, B)\}\rangle$. Clearly, here the preferred extension $E = \{A, C\}$.

**Example 2.** Let $AF = \langle\{A, B\}, \{(A, B), (B, A)\}\rangle$. There are two preferred extensions, $\{A\}$ and $\{B\}$.

**Example 3.** Let $AF = \langle\{A, A\}, \{(A, A)\}\rangle$. Here the preferred extension is the empty set.

**Definition 12 (Stable Extension).** A conflict-free set of arguments $S$ will be a stable extension (SE) iff $S = \{A|\forall B \not\in S \text{ will be attacked by } S\}$. 

44
In examples 1 and 2 above, the preferred extensions are also stable extensions. But in example 3, the empty set is not a stable extension.

The preferred and stable extensions are considered to both be credulous — they consider an argument to be acceptable when a more skeptical approach might not. Argumentation also has more skeptical notions of an extension which we will introduce below.

**Definition 13 (Characteristic Function).** Dung defined a monotonic characteristic function $F_{AF}$ that returns the acceptable sets for each input set.

That is,

$$F_{AF} : 2^{AR} \rightarrow 2^{AR}$$

$$F_{AF}(S) : \{ A | A \text{ is acceptable with respect to } S \}$$

Dung also showed that if the argumentation framework is finitary which is, for each argument, there are a finite number of arguments that attack it, then $F_{AF}$ is continuous and its least fixed point can be can be found by iteratively applying it to the empty set.

**Definition 14 (Complete Extension).** An admissible set $S$ is a Complete Extension iff all arguments defended by $S$ are also in $S$.

There could be more than one complete extension each corresponding to a particular viewpoint.

**Definition 15 (Grounded Extension).** A conflict-free set of arguments $S$ is the Grounded Extension if it is the minimal (with respect to set inclusion) complete extension.

defeated as well as all those arguments that are supported directly or indirectly by these un-attacked arguments. A grounded extension is also the Least Fixed Point of $F_{AF}$.

In example 1, $\{ A, C \}$ is also the grounded extension, but in example 2, the grounded extension is empty. In other words, we can say that a skeptical reasoner will conclude nothing if the grounded extension is empty.

### 3.2 First Scenario: Target Has No Historical Data

In this scenario, we consider the situation when an agent needs to transact with another one with whom he has no previous experience and no referral. Suppose, a buyer $X_i$ has to buy a product from seller $Y_j$. Neither $X_i$ nor any of his neighbors have any previous transactions with $Y_j$. How is $X_i$ going to decide if he will trust $Y_j$ or not? Consider the conversation between $X_i$ and $Y_j$ to be as follows:

- $X_i$'s claims: $\{ A, B, C, D \}$
- $Y_j$'s claims: $\{ a, b, c, d \}$

From Figure 1(a), we can see that $Y_j$’s claim $a$ is attacked by $X_i$’s $A$. $Y_j$ backed up his claim with $b$, which attacks $A$. $b$ is again attacked by $X_i$’s claims $C$ and $D$. In the same way, $D$ is attacked by $d$ and $c$, and $B$ attacks $c$. At this point, $X_i$ will build an argumentation framework $AF$ out of the conversation and compute the stable extensions. Here we are assuming a meaningful argument should be well-founded and coherent. As we
are also assuming all agents are skeptical reasoners, they will decide nothing unless a winning extension is found.

$X_i$’s $AF$ and the generated stable extensions will look like the following (we follow Dung’s notion of abstract argumentation here):

$$F = \langle \text{Arg}, \text{Att} \rangle$$

$$= \langle \{A, B, C, D, a, b, c, d\}, (A, a), (b, A), (C, b), (D, b), (d, D), (c, D), (B, c) \rangle$$

where $SE_{X_i} = \{A, B, C, D\}$ and $SE_{Y_j} = \{a, b, c, d\}$.

As our agents are skeptical reasoners, they tend to follow the grounded extension as its conclusions are not controversial. Figure 1(a) shows, $X_i$ has two unattacked claims $B$ and $C$ and $Y_j$ has one unattacked claim $d$. At this point, we can say $X_i$ has two arguments that no one can attack but can we say that $X_i$ has two arguments that no one can disprove? No. If we follow this process, anyone in $Y_j$’s place can provide as many arguments as he can for the sake of winning. Unfortunately, this could happen both ways around and could go on and on, which will destroy the well-founded structure of the framework.

Instead of computing the conventional grounded extension (GE), we propose an extendedGE, which is limited to considering evidence as the starting point. In our model, evidence is the vital element to win an argument. Our opinion is that, if someone is saying something true he should be able to support his claim with evidence. Here by “evidence”, we are indicating statements about the ground truth of the domain which are non-conflicting if the domain is consistent. At this point, neither $X_i$ nor $Y_j$ has provided any evidence. This forces them to supply evidence to fortify their claims. In Figure 1(b), we see that, $X_i$ supports his claim $C$ with evidence $r$. As evidence provides the non-conflicting ground truth, the attacked arguments will automatically be eliminated. Therefore, $r$ eliminates $b$. Likewise, $Y_j$ supports his claim $c$ with evidence $q$, which eliminates $B$. The agent must iteratively return arguments which are themselves evidence or have evidence as a supporting argument. These will eliminate the arguments.
that these arguments attack. This process follows until the agents reach a conclusion. After elimination, the new stable extensions will look like the following:

\[
SE_{X_i} = \{A, C, r\}\\
SE_{Y_j} = \{a, c, d, q\}
\]

At this point, both of the stable extensions have exactly one piece of evidence. In our model, the evaluator will break this tie by considering the depth of the supporting evidence. We extend the idea of depthLimitR from Yu and Singh [22] and propose depthLimitE which denotes the number of hops the evidence is away from the claim it is supporting. In the example above, \(r\) is two hops away from the initial claim \(A\), and \(q\) is six hops away from its initial claim \(a\). Intuitively, evidence is more relevant if \(depthLimitE\) is short, and evidence becomes more irrelevant as \(depthLimitE\) increases. This makes \(SE_{X_i}\) the winner. Therefore, \(Y_j\) fails to defend his claims, and \(X_i\) will rate \(Y_j\) from its lower trust limit range which will be used later to compute the BPA of \(Y_j\), and afterwards belief in \(Y_j\).

\[
\begin{align*}
\{\lnot T\} & \quad \omega=0.4 \\
\{T, \lnot T\} & \quad \Omega=0.8 \\
\{T\} & \quad 1
\end{align*}
\]

Fig. 2. Trust Scale of agent \(X_i\)

Consider the following example: \(X_i\)’s upper trust limit is \(\Omega = 0.8\) and lower trust limit is \(\omega = 0.4\). Therefore, all the transactions with \(\rho = [0.8, 1.0]\) count for \(\{T\}\) and \(\rho = [0, 0.4]\) count for \(\{\lnot T\}\). The rest count for \(\{T, \lnot T\}\). The scenario we present is a special case where \(X_i\) has no previous data about \(Y_j\). \(X_i\) will select a value from its lower trust limit range \([0, 0.4]\), depending on how badly \(Y_j\) failed to defend himself as a prior rating for \(Y_j\). Let \(X_i\) select 0.2 as the initial rating for \(Y_j\) and \(X_i\)’s probability of making a good decision as 0.8. A potential way of measuring initial local belief \(Bel\) in \(Y_j\) could be:

\[
Bel(Y_j) = \frac{\text{Good decisions taken by } X_i}{\text{Total decisions taken}} \times \text{Prior rating for } Y_j = 0.8 \times 0.2 = 0.16
\]

Though \(X_i\)’s probability of making a good decision is high, the result is low due to \(Y_j\)’s poor rating. If this value is below the risk threshold, then \(X_i\) will not engage in any communication or transactions with \(Y_j\).

The idea of depthLimitE to count the number of hops across pieces of evidence could later be used in risk analysis. As we said before, evidence is more relevant when it supports claims closer to the primary claim. Hence, we can say:

\[depthLimitE \propto \text{risk}\]
Some researchers propose a semantics (the ideal semantics) that is less skeptical than grounded extension but more skeptical than preferred extension [6]. In our model, we can control the skepticism by taking \( \text{depthLimit}_E \) into account. Intuitively:

\[
\text{depthLimit}_E \propto \frac{1}{\text{skepticism}}
\]

We assume that each agent has its distinct risk threshold which solely depends on the current state of that agent. A higher risk threshold indicates the agent is capable of taking more risk. Therefore we can say that an agent with a high risk threshold can choose to consider evidence with larger \( \text{depthLimit}_E \). Thus \( \text{depthLimit}_E \) could be a potential factor to consider in analyzing trust sensitivity. Note that we reserve discussion of risk analysis and trust sensitivity for future work.

### 3.3 Second Scenario: Target has Transaction History

In this section, the seller is known to the buyer. As the buyer has had previous transactions with the seller, it will build its local trust from the previous trust rating using Dempster-Shafer theory. Consider the following: \( X_i \) has had six previous transactions with \( Y_j \). After the last transaction, \( Y_j \)'s ratings are, say, \( \{0.2, 0.6, 0.9, 0.7, 0.3, 0.2\} \). Let \( x \in \{0.2, 0.6, 0.9, 0.7, 0.3, 0.2\} \). According to Dempster-Shafer theory, \( Y_j \)'s BPA will be:

\[
\begin{align*}
m(T) & = \sum_{Q=0.8}^{1} f(x) = 1/6 \times 3 = 0.167 \\
m(\neg T) & = \sum_{Q=0.4}^{0} f(x) = 1/6 \times 2 = 0.333 \\
m(T, \neg T) & = \sum_{Q=0.8}^{0} f(x) = 1/6 \times 3 = 0.5 \\
\end{align*}
\]

Therefore, the belief values for \( Y_j \) would be:

\[
\begin{align*}
\text{Bel}(T) & = 0.167 \\
\text{Bel}(\neg T) & = 0.5 \\
\text{Bel}(T, \neg T) & = 0.333 \\
\end{align*}
\]

As we can see, \( \text{Bel}(T) - \text{Bel}(\neg T) \) is negative (-0.333), which is obviously a lot less than the trust threshold. Thus the buyer will not engage in any transactions with the seller.

In the equation used for deciding “to trust” or “not to trust”:

\[
\text{Bel}(T) - \text{Bel}(\neg T) \geq \text{trust threshold}
\]

if the difference is large (the seller is either highly trusted or highly distrusted), then it will follow the same process. But if the difference is small, which is, a big number of transactions fall under an uncertain state, then the buyer will follow the process of the first scenario, outlined in Section 3.2, to see if the current transaction can limit the uncertain state.
3.4 Third Scenario: Combining Trust in a Prebuilt Trust Network

In most practical cases, the evaluator or buyer does not have enough transactions or has no transactions at all with the desired seller. Here, buyers often look for referrals to learn something about the seller. The situation where no referrals are available was described in Section 3.2. Now we are going to describe the case of combining referrals. In our model, the buyer or evaluator sends out the query to its trusted neighbors asking for testimonies about the seller. If the neighbors have past experience and have built a local belief structure (in the way described in Section 3.2), then they pass their belief value(s) to the buyer. In cases where the seller is also unknown to the neighbor, the neighbor may pass a referral on to a potential agent who may have past experience with the seller. This process follows until the evaluator gets the desired testimony (or there are no more agents left to query). As mentioned earlier, in [22], the authors present depthLimitR, which denotes the length of the referral chain. We introduce some additional constraints here. If depthLimitR falls outside of a given range, then the seller will be treated as a newcomer with no referral history; and the scenario described in Section 3.2 will be followed. This range will be set by risk analysis, which is a topic we reserve for future work.

Fig. 3. Local trust propagation in pre-built trust network

4 Multiple belief values may exist, for example, where beliefs are contextualized and a vector associates individual beliefs with a set of contexts. Here, we abstract the notion of belief into a single value and reserve discussion of belief as a complex data structure for future work.
Consider the graph in Figure 3. Our buyer, $A$, sends out a query about seller $C$ to its trusted neighbors $B$, $D$, and $E$. Among them, only $B$ has previous experience with $C$ and has hence built a local belief structure about $C$. This local belief will be passed on as a testimony to $A$. The transaction will be between $A$ and $C$; and $C$’s dialogues, along with its testimony, will be passed to $A$. At the same time, $D$ and $E$ will pass the query to $G$ and $F$, respectively. It is a very common scenario in practical cases that $G$ and $F$ have no information about $C$, but they do have experience about the product he is selling—which is crucial in making a decision, but was not explicitly requested in the query. These claims, along with the testimonies, will be passed to $A$ in a similar fashion.

Local belief values will be merged using the method proposed in [21]. We use the concatenation and aggregation operators proposed in [13], and subsequently used by [21], to merge the trust values in the graph. The concatenation operator is used to merge trust within the same referral chain. On the other hand, the aggregation operator is used to combine the trust values on the same topic that come from different sources (agents). In our example, consider that $A$’s local belief towards its trusted neighbors $B$, $D$, and $E$ are $M_B$, $M_D$, and $M_E$, respectively. Again, $B$ holds $M_C$, its local belief structure concerning $C$, $E$ holds $M_{DF}$, its beliefs in $F$, and $D$ holds $M_G$, its beliefs in $G$. Here, $M_B = Bel_B$. This belief function has three parts: “Belief” in $B$, “Disbelief” in $B$ and “Uncertainty” about $B$. Separately, $Bel(T) = m(T)$, $Bel(\neg T) = m(\neg T)$ and $Bel([T, \neg T]) = m(T, \neg T)$, which are the summation of the probabilities of “Good transactions”, “Bad transactions” and “Uncertainty”, respectively (as discussed above).

For simplicity and similarity, we will follow the notions used in [21]. Let,

$$m(T) = P_B$$
$$m(\neg T) = N_B$$
$$m([T, \neg T]) = U_B$$

Therefore, following [21], we construct $A$’s primary beliefs about $C$ as follows:

$$M_{AC} = M_B \otimes M_C$$
$$P_{AC} = P_B \times P_C$$
$$N_{AC} = P_B \times N_C$$
$$U_{AC} = 1 - (P_B \times P_C) - (P_B \times N_C)$$

Here, $\otimes$ is the concatenation operator, which is just Dempster’s rule from before. At this point, we can say that $A$ has $M_{AC}$ primary belief in $C$’s claim $[a, b]$. $M_{AF}$ and $M_{AG}$ will be constructed in a similar way. We mention $A$’s “primary belief” in $C$ because $A$ still has to assimilate all of the information he gets from $G$ and $F$ to come up with his final belief.

Next, all these dialogues from $C$, $F$ and $G$ will be put in an argumentation framework, along with $A$’s knowledge similar to the scenario described in Section 3.2, except that the pivotal point will be the “combined local belief” in those claims. That means $A$ will consider the following in constructing the argumentation framework:

$$M_{AC}[a, b] \cup M_{AG}[i] \cup M_{AF}[p, q] \cup DA$$
Here, $D_A$ is $A$’s knowledge about the domain. Following our earlier assumption, high-valued claims will be prioritized over low-valued claims, while defeating each other. If there is a tie (same combined trust), then the scenario in which there is no prior history (Section 3.2) will be followed again, and this time “evidence” will be used as the tie-breaker.

3.5 Discussion

This section highlights three issues that have not been specifically addressed above, but need to be considered when using argumentation to compute trust. These issues are: the Fake Profile problem, the Trust Transitivity problem, and the Incentive problem. Each is discussed briefly, below.

The Fake Profile problem is a major issue in reputation systems. Membership in most social networking and business rating sites such as Yelp$^5$, for example, is free. As a result, there is very little to stop people creating many different profiles with which they boost or downgrade the reputation of an entity. These fake profiles have a very bad impact on cooperation or even initiation of a transaction. This impacts how newcomers will be treated [8]. In our model, every agent has to defend his claims with evidence. No matter how many profiles that agent has received or how many good referrals were collected, in the end, he needs to hold evidence. This requirement suppresses fake profiles to a great extent. Moreover, as shown in [18], with enough exchanges of arguments, it is not possible for one agent to deceive another indefinitely—eventually their knowledge bases converge.

In belief systems, Trust Transitivity is another major issue. It is possible that what the evaluator decides is most heavily influenced by its witnesses’ beliefs. In this case, making decisions that depend upon witnesses’ local beliefs is prone to deception. There are several proposals in the literature addressing trust transitivity [3, 10, 14, 19, 22]. In our model, since contributing agents are invisible to each other (e.g., in Figure 3, $C$, $F$ and $G$ are invisible to each other), a malicious agent does not gain any advantage by deceiving a trusted node, as he does not necessarily know who opposes his claim. This leaves him with no choice but to deceive large numbers of agents, possibly all of an evaluator’s trusted nodes! And, in the end, the deceiver is required to show evidence; trust transitivity does not help him much here.

There are cases when trusted agents exist but have little incentive to contribute information to third-party transactions. To encourage them to participate, we propose $\text{rewardVal}$ and $\text{penaltyVal}$, respectively. The latter, $\text{penaltyVal}$ will decrease the agent’s rating and hence belief in him. Similarly, the former, $\text{rewardVal}$, will increase this rating. This will incentivize agents to contribute and will penalize malicious agents for infusing unauthentic information. The risk threshold of the agent can thus be compared to the $\text{penaltyVal}$ and $\text{rewardVal}$ to optimize decision making. If these values are made public, then it is possible to guess an agent’s current state by analyzing these values. Moreover, if an agent is willing to deceive and can afford the $\text{penaltyVal}$ (i.e., $\text{penaltyVal} < \text{riskThreshold}$), then he may take the risk of deceiving the evaluator agent. We will discuss these values more in our future work.

$^5$ http://www.yelp.com
4 Summary and Future Work

In our model, we have addressed some problems in current trust management and reputation systems by incorporating evidence into an argumentation framework, and then integrating it into multiple trust management scenarios. In future work, our plan is to refine the theory and focus more on risk analysis. In particular, we are considering adding the concept of utility to our trust management models in order to capture the differential importance of evidence to different agents. This might be used to perform a risk analysis to judge the effects of making incorrect trust-based judgments. We also intend to investigate foundations and formulations for assigning trust thresholds and choosing ratings to measure BPA which will make our model more precise. Later, we plan to implement it in a more practical environment, such as a recommendation system for online social network applications.

5 Acknowledgments

Research was partially funded by the National Science Foundation, under grant CNS 1117761 and Army Research Laboratory and Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, the National Science Foundation, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

A Formal Argumentation Dialogue for Personalised Trust Communication

Andrew Koster\textsuperscript{1,2}, Jordi Sabater-Mir\textsuperscript{1}, and Marco Schorlemmer\textsuperscript{1}

\textsuperscript{1} IIIA-CSIC
Bellaterra, Spain
\{andrew, jsabater, marco\}@iiiia.csic.es
\textsuperscript{2} Universitat Autònoma de Barcelona
Bellaterra, Spain

Abstract. In order to allow agents in a multi-agent systems to effectively communicate about trust, we propose to personalise the communication about trust. In previous work we described AdapTrust and an argumentation framework; the former to adapt the trust model to the needs of the receiving agent and the latter to formalise a language in which to communicate. In this paper we propose a dialogue protocol for this communication, formalising it as an extension of Prakken’s dialogue game for an information-seeking dialogue.

1 Introduction

Trust models for computational agents often center around the processing of evidence from direct interactions with a target. Nevertheless, in sufficiently large multi-agent systems it is unfeasible for agents to have had previous direct interactions with all targets they may need to interact with. It is thus essential that they can communicate about their trust evaluations, in order to decide whether or not a target is trustworthy. Trust, however, is a personal and subjective evaluation of a target for the fulfillment of a specific goal, and it is not straightforward to communicate this.

So far, communication methods for trust have relied mainly on machine-learning techniques to translate another agent’s evaluation into the receiver’s own frame of reference [1,5,12]. All of these methods learn a translation based on interactions that the recommendation-seeking agent shares with the recommendation-supplier, and for this to succeed the agents must share a large number of interactions. If this is not the case, the agents cannot learn an accurate translation and another method is required.

In previous work we proposed that the recommendation-supplier adapts, or personalises, its trust evaluation to the specific needs of the recommendation-seeker, rather than the seeker attempting to translate the unadapted recommendation from the supplier [4]. We presented an argumentation framework for communicating the beliefs that influence the agent’s trust evaluation and the goal for which it evaluates the target. In this paper we propose a formal dialogue protocol for this argumentation. This dialogue protocol clearly defines the
moves an agent is allowed to make when communicating in order to personalise communication about trust.

In Section 3 we present a dialogue protocol for the argumentation and personalisation of trust, but first we briefly summarise our argumentation framework and a prerequisite for it: a method for incorporating the trust model into an intelligent agent. AdapTrust is an extension to the BDI-agent model that allows an agent to adapt its trust model according to its beliefs and goals.

2 Preliminaries

Our method for enabling personalised communication about trust is based on three capabilities an agent must have. Firstly, an agent must be able to adapt its trust model in order to personalise its evaluations to the other agent’s needs. Secondly, an agent must be capable of communicating its criteria for evaluating trust, as well as the underlying beliefs and goals leading to these criteria. Finally, an agent must be willing and able to change its trust model, if it is persuaded that its beliefs about the environment, and thus the criteria for calculating trust are wrong.

We assume that agents are willing to adapt their model if they are convinced it is inaccurate. For this adaptation to be possible, the trust model must be incorporated into the cognitive reasoning process of the agent, and we propose to use AdapTrust for this [6]. To communicate about how the trust model needs adapting in order to personalise a trust recommendation we use an argumentation language [4] and the dialogue protocol that we present in this paper. We start with a brief explanation of AdapTrust and the argumentation language.

2.1 AdapTrust

Computational trust models are, fundamentally, methods of aggregation: they combine and merge data from several different sources into a single value, the trustworthiness of a target. Moreover, this computation depends on the beliefs the evaluator has about the world, as well as the goal it is trying to achieve. Luckily most computational trust models come equipped with a way of implementing this dependency: they have parameters that can be used to adjust the behaviour of the trust model. The aim of AdapTrust is not to present another trust model, but to incorporate existing trust models into an intelligent agent [6]. AdapTrust works by changing the parameters of the trust model in accordance with the beliefs an agent has about the environment and the goal the agent wants to achieve, and for which it needs the trust evaluation.

Priority System The parameters of a trust model describe the importance of the different criteria for evaluating trustworthiness. However, it is more useful to consider this the other way round: the relative importance between the different criteria define a set of parameters for the trust model. These criteria are directly under an intelligent agent’s control, and thus an agent is able to adapt its trust model. AdapTrust describes the specific techniques necessary to do this. The first of these is $\mathcal{L}_{PL}$, a language to describe the relative importance of any
two criteria that influence a parameter of the trust model. We use a subset of first-order logic, with a family of special predicates to define this importance relation, also called a priority ordering. For each parameter $p$ of the trust model, the binary predicates $\succ_p$ and $\equiv_p$ are defined with the expected properties of strict ordering and equality, respectively. The language uses a set of constants to represent the criteria that influence how the trust model should work. A Priority System is defined as a satisfiable theory in this language. For instance, consider an eCommerce environment. If an agent uses a weight $w$ to calculate its evaluation of a sale, and it finds the price of an item to be more important than its delivery time, it can have the priority $\text{price} \succ_w \text{delivery.time}$ in its Priority System.

**Priority Rules** The second technique of AdapTrust is to create the link between, on the one hand, an agent’s beliefs and goals, and on the other hand, the priority between the different criteria for evaluating trust. This link makes explicit the adaptive process: a change in an agent’s beliefs or goals affects a change in the priorities over the criteria, which in turn changes the parameters of the trust model. The connection between the beliefs or goals and the priorities is made through what we call priority rules. The priority rules are specified using another first-order language, $\mathcal{L}_{\text{Rules}}$, with predicates $\succ_{\text{Belief}}$ and $\succ_{\text{Goal}}$ specifying how a set of beliefs, or a goal, respectively, leads to a specific priority relation between two criteria. By using these rules, we see that when the belief base changes the priorities can change. Additionally this is how the multifaceted aspect of trust is emphasised: the goal the agent is trying to achieve influences the priority system and thus the trust model. For instance, in an eCommerce example, an agent might need to buy an item urgently. It then has the goal $\text{buy.urgent(item)}$. For this goal, delivery time is more important than the price, so it has the priority rule $\text{buy.urgent(item)} \succ_{\text{Goal}} (\text{delivery.time} \succ_w \text{price})$. This adapts its trust model to the requirements of the goal. The priority rules are atomic predicates, rather than implications in $\mathcal{L}_{\text{Rules}}$, because standard first-order semantics for material implication should not hold. For instance, if the agent does not have $\text{delivery.time} \succ_w \text{price}$ in its priority system, this does not mean that the agent does not have the goal $\text{buy.urgent(item)}$. There may just be other, conflicting priority rules that have precedence over this rule.

### 2.2 Argumentation Framework

Our argumentation framework extends Pinyol’s framework for arguing about trust [8], and is explained in more detail in [4]. In this section we summarise the argumentation framework and language.

The first requirement for arguing about trust is that the agents have a common language in which to describe their trust evaluations. We use the $\mathcal{L}_{\text{Rep}}$ language, described by Pinyol et al. [9]. This is a first-order language about trust and reputation, defined by a taxonomy of predicates used for describing the process of computing trust. Some of the predicates describe what they refer to as ground elements, such as direct experiences and communications. We represent the ground elements as the set $\text{ground}(\mathcal{L}_{\text{Rep}})$. Other predicates describe
“higher” concepts, such as the outcome of a direct experience or the reputation of a target. In Pinyol’s argumentation framework, $L_{Rep}$ is sufficiently expressive for all the communication, but we need to extend the language we use. First, however, we describe one thing our frameworks have in common: the interpretation of a trust model as an inference relation.

**Trust Models and Inference** A key point of both Pinyol’s argumentation framework and our own is the focus on how to generate arguments. For this we build on the representation of any computational process as the application of a finite set of inference rules [3]. A trust model is a computational process and can thus be represented by a set of inference rules. The process of calculating a trust evaluation can be seen as the finite application of a number of inference rules $I$ on a set of inputs $\Delta \subseteq \text{ground}(L_{Rep})$ to obtain the output $\delta \in L_{Rep}$. We write $\Delta \vdash_ I \delta$. The inference rules themselves depend on the specifics of the computational process and thus the actual trust model being used, but for any computational trust model, such an inference relation exists. An example could be to infer the trust evaluation from reputation as follows:

$$\frac{\text{rep}(T, X)}{\text{trust}(T, X)}$$

The main difference between Pinyol’s framework and our own, is that we assume the trust model is integrated into the agent’s cognitive process by using AdapTrust, and it is therefore dependent on the agent’s beliefs and goals: trust is an evaluation of a target for a specific goal, given the evaluator’s beliefs about the environment. These beliefs and this goal influence how the trust model computes an evaluation and this must be represented as well in the inference rules. We assume the agent’s beliefs and goals are represented in logical languages $L_{Bel}$ and $L_{Goal}$, as is the case in AdapTrust. For a set of beliefs $\Psi \subseteq L_{Bel}$ and a goal $\gamma \in L_{Goal}$ we have a set of inference rules $I^{\Psi, \gamma}$, and we write $\Delta \vdash_ {I^{\Psi, \gamma}} \delta$ to represent that input $\Delta \subseteq \text{ground}(L_{Rep})$ results in trust evaluation $\delta \in L_{Rep}$ for goal $\gamma$, given beliefs $\Psi$.

The way these inference rules are affected by the beliefs and goal is defined in AdapTrust: a set of beliefs and a goal cause certain priority rules to trigger, which leads to a set of priorities. A set of priorities describes a legal set of values for the parameters and in this way the trust model is adapted to the beliefs and goals. Not all inference rules are affected by the same priorities, because not all inference rules use the same parameters. We thus see that for a set of beliefs $\Psi$ and a goal $\gamma$, we have that for any $\iota \in I^{\Psi, \gamma}$ there is a (possibly empty) set of parameters $\text{params}(\iota)$. The values for the parameters, in turn, are prescribed by a set of priorities $P^{\Psi, \gamma}$.

**Arguing about Trust** To be able to communicate about the trust process we must describe a formal language. We use the argumentation framework presented by Chesniev and Simari, $LDS_{ar}$ [2], which provides an intuitive way for representing the inference rules $I$ in a communication language $L_{Arg}$. $L_{Arg}$ is a labelled language for defeasible reasoning, but for simplicity we omit the labels (for the full formalisation see [4]). We interpret it as a non-monotonic
propositional language, in which we allow the connectives $\land$ as conjunction, and $\rightarrow$ as non-monotonic implication with semantics as in logic programming (for the formal semantics, see [2]). The language has three deduction rules, which are:

\begin{align*}
\text{Intro-BDU: } & \quad \alpha \quad \text{for any } \alpha \in \mathcal{L}_{KR} \\
\text{Intro-AND: } & \quad \alpha_1, \ldots, \alpha_n \quad \alpha_1 \land \cdots \land \alpha_n \\
\text{Elim-IMP: } & \quad \alpha_1 \land \cdots \land \alpha_n \rightarrow \beta, \alpha_1 \land \cdots \land \alpha_n
\end{align*}

These deduction rules are used to deduce the conclusion of an argument from the argumentative theory, which is a set of basic sentences in $\mathcal{L}_{Arg}$ that are called basic declarative units (bdus). These bdus are ground sentences in an underlying language for knowledge representation $\mathcal{L}_{KR}$. In Pinyol’s framework, but we will extend this.

Let the agent have beliefs $\Psi$ and goal $\gamma$ for which a trust model infers trust evaluation $\delta$ from input $\Delta$. We write $\Delta \vdash_{\Psi, \gamma} \delta$ using inference rules $\mathcal{T}^{\delta, \gamma}$. Let $\iota \in \mathcal{T}^{\delta, \gamma}$ be an inference rule such that $\alpha_1, \ldots, \alpha_n \vdash \beta$, with $\alpha_1, \ldots, \alpha_n, \beta \in \mathcal{L}_{Rep}$, and the values of the parameters $\text{params}(\iota)$ are prescribed by the priorities $\Pi_{\Psi, \gamma}$; then we add a bdu $(\land_{\pi \in \Pi_{\Psi, \gamma}} \pi) \rightarrow (\alpha_1 \land \cdots \land \alpha_n \rightarrow \beta)$ in $\mathcal{L}_{Arg}$. We do this for all $\iota \in \mathcal{T}^{\delta, \gamma}$.

Furthermore we add a bdu for each priority rule: if $\Phi \sim_{\text{Belief}} \pi$ is a priority rule, then $\Phi \rightarrow \pi$ is a bdu. The same for any goal $\gamma' \sim_{\text{Goal}} \pi$, we have $\gamma' \rightarrow \pi \in \mathcal{L}_{Arg}$. We also add all $\delta' \in \Delta$, all the agent’s beliefs $\psi \in \Psi$ and the agent’s goal $\gamma$ as bdus. This means that the knowledge representation language $\mathcal{L}_{KR}$ that underlies $\mathcal{L}_{Arg}$ must be extended too. We have $\mathcal{L}_{KR} = \mathcal{L}_{Rep} \cup \mathcal{L}_{PL} \cup \mathcal{L}_{Rules} \cup \mathcal{L}_{Bel} \cup \mathcal{L}_{Goal}$.

The set of bdus generated in this manner gives a way for an agent to justify its trust evaluation in $\mathcal{L}_{Arg}$. While the above description seems to imply that an agent starts from the ground elements in $\mathcal{L}_{Rep}$, its beliefs and its goals, to generate bdus in $\mathcal{L}_{Arg}$, in actual fact the reverse is true. An agent uses its trust model, influenced by its beliefs and goal, to calculate a trust evaluation. It then traces this process back and encodes its calculation process, and the inputs it used, as bdus in $\mathcal{L}_{Arg}$. Because sentences in $\mathcal{L}_{Arg}$ are communicable between agents, any agent can follow another’s reasoning and deduce the trust evaluation from the inputs without knowing any of the details of the other agent’s trust model. By following this deduction process an agent can also reason about whether it agrees, or disagrees, with the other agent and why. We note the main reasons agents may disagree about the trust model:

- The agents disagree about (some of) the ground elements $\Delta \subseteq \mathcal{L}_{Rep}$ that are introduced into $\mathcal{L}_{Arg}$ as bdus. This was dealt with in Pinyol’s argumentation framework and we do not go into detail about that [8]. In general we will assume that if agents disagree about the ground elements, then communication fails, but it should not happen often. In general the recommendation-seeking agent is asking for advice about a target that it has no, or little, knowledge
of, and it can accept that the recommendation-supplier has had a number of direct experiences with the target.

- The agents disagree about (some of) the beliefs $\Psi \subseteq \mathcal{L}_{Bel}$ that are introduced as bdus. In this case the agents can enter a persuasion dialogue to try to reach an agreement about beliefs.

- The agents have different goals. The recommendation-seeking agent should make it clear from the start that the recommendation is needed for a specific goal, and the recommendation-supplier should use this goal in its trust computation. If this does not happen, the recommendation-seeker should reject the recommendation.

- The agents disagree about a set of priorities that beliefs $\Psi$ and goal $\gamma$ lead to. In this case, the agents have different priority rules in AdapTrust. They can communicate these priority rules between each other and adapt their trust model.

- Despite having the same priorities, the agents disagree on the evaluation that can be inferred from a set sentences $\alpha_1, \ldots, \alpha_n \in \mathcal{L}_{Rep}$. In this case the agents’ computational process is too different to be able to adapt: they agree on all the premises, but not on the conclusion. The recommendation-seeker should reject the recommendation from that supplier and try to communicate with a different agent.

In the next section we present a formal dialogue protocol in which two agents can argue about a trust evaluation and find places that they disagree. It then allows them to deal with this disagreement in the way we described above.

3 Dialogue Protocol for Personalising Trust

The argumentation in the previous section can be used by an individual agent to justify its trust evaluation in a language that the other agents understand. We now specify a protocol that allows agents to argue back and forth in order for the requesting agent to receive a personalised trust recommendation from the witness. We illustrate this protocol with an example, and start with explaining this example.

3.1 An example of argumentation

An argument for a trust evaluation can be represented in a tree. We call this an argumentation tree and give an example of one in Figure 1. The argumentation tree can be followed by applying the deduction rules of $\mathcal{L}_{Arg}$ at each level. In order to be succinct, we use shorthand in the tree by referring to nodes, rather than repeating the content of a node. For instance, in node $R_1$ we can expand $E_2 \land E_3 \rightarrow E_1$ to its meaning: $img(Jim,5) \land rep(Jim,1) \rightarrow trust(Jim,5)$, where $img$ and $rep$ are predicate symbols in $\mathcal{L}_{Rep}$, and are short for the agent’s image, and reputation of the target. An argumentation tree, such as this one, is used in a dialogue to communicate personalised trust evaluations.

We do not explore all the paths in the tree and leave the nodes $E_2$ and $E_3$ unexplored, because their unfolding results in a similar structure to the unfolding of the root ($E_1$).
3.2 A formal dialogue protocol

We can now define a formal dialogue system for communication about personalised trust recommendations in which the argumentation can be communicated. The system we need is, for a large part, an information-seeking dialogue system, according to the classification by Walton and Krabbe [13]. It thus stands to reason that we use a protocol similar to the one presented by Parsons et al. [7]. However, while our dialogue is for a large part information-seeking, it also incorporates some aspects of persuasion dialogues. We thus present the formal system in a similar structure to the dialogue system presented by Prakken [10] for persuasion dialogues, in order to allow for some locutions in addition to the “question”, “assert” and “challenge” locutions proposed by Parsons et al.

Definition 1 (Dialogue System for Personalised Trust (adapted from Prakken’s Definition 3 [10])). A dialogue system for personalised trust is a tuple $\mathcal{D} = (\mathcal{L}_C, P, CR)$ where $\mathcal{L}_C$ (the communication language) is a set of locutions, $P$ is a protocol for $\mathcal{L}_C$, and $CR$ is a set of effect rules of locutions in $\mathcal{L}_C$, specifying the effects of the locutions on the participants’ commitments.

The three parts are described below, but first we must define some of the basic elements of a dialogue. The first of these is the set of participants themselves. These participants of the dialogue are the recommendation-seeker and recommendation-supplier, and we denote them with $Q$ and $R$, respectively. Both of these agents have a commitment store, a set of sentences in $\mathcal{L}_{Arg}$ that they have committed themselves to [13]. Commitment is a complicated concept, but we use it in a very specific way: an agent’s commitment store contains beliefs it has voiced during the dialogue and is committed to justify and defend. Because the dialogue is essentially an information-seeking dialogue, the recommendation-supplying agent $R$ will mainly be the one committing itself to sentences in the dialogue. As the dialogue progresses, the recommendation-supplier will justify, in increasing detail, why the initially communicated trust evaluation holds. Every justification of this kind adds to the recommendation-supplier’s commitment store. The agents’ commitment stores are denoted $C_Q$ and $C_R$ for agents $Q$ and $R$, respectively. Initially both agents’ commitment stores are empty.
With these concepts in place we can move on to the definition of the locutions and protocol of a dialogue system. We start with the locutions.

**Definition 2 (Locutions for Personalised trust).** The locutions allowed in the dialogue for personalised trust are specified by $L_C$ and include the basic locutions for information-seeking, specified by Parsons et al. [7]. The locutions are explained in Table 1.

Some of the locutions have an effect on an agent’s commitment store. We usually denote the agent (either $Q$ or $R$) that is sending a message, also called making a move, with $s$ and the other agent with $\overline{s}$. We take $C_s'$ to be the new commitment store of agent $s$ after sending the location, and $C_s$ is the old commitment store prior to sending. The way the commitment store is updated for each location is detailed in Table 2, which thus defines the rules $CR$ of the dialogue.

<table>
<thead>
<tr>
<th>Location</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>request_recommendation($t, \gamma$)</td>
<td>The initial request for a recommendation, with $t \in Agents$ and $\gamma$ the goal that $Q$ wants the recommendation for.</td>
</tr>
<tr>
<td>assert($p$)</td>
<td>Assert that $p$ is true, where $p \in L_{Arg}$.</td>
</tr>
<tr>
<td>justify($p, S$)</td>
<td>Assert that $S \subseteq L_{Arg}$ is the (direct) support for $p$ in $L_{Arg}$.</td>
</tr>
<tr>
<td>challenge($p$)</td>
<td>Challenge a sentence $p \in L_{Arg}$ in the other agent’s commitment store. An agent may challenge a sentence $p$ if it wants the other agent to justify $p$.</td>
</tr>
<tr>
<td>counter($\pi_R, \pi_Q$)</td>
<td>Propose an alternative priority $\pi_Q$ to priority $\pi_R$ with $\pi_Q, \pi_R \in L_{PL}$. Note that this switches roles: counter is similar in use to assert, so the agent $Q$, that has thus far only been challenging assertions, now proposes its own priority, that $R$ can now challenge.</td>
</tr>
<tr>
<td>argue($\psi$)</td>
<td>Propose to enter into a separate persuasion dialogue about beliefs $\psi \subseteq L_{Bel}$. The details of this dialogue are outside the scope of this paper, but we propose to use the dialogue system proposed by Prakken [11].</td>
</tr>
<tr>
<td>end</td>
<td>Indicate that the dialogue has concluded.</td>
</tr>
</tbody>
</table>

Table 1. Locutions in $L_C$, the communication language for personalised trust recommendation dialogues

The locutions request_recommendation, assert and challenge correspond directly to “question”, “assert” and “challenge” in Parsons et al.’s system. Moreover, justify also corresponds to “assert” in Parsons et al.’s framework, but because they do not allow agents to backtrack, the sentence being justified is always immediately clear from the previous dialogue steps. The locutions counter and argue are not present in regular information-seeking dialogues. We add these so that agents can propose alternative priority systems for AdapTrust.
Table 2. The effect of the various locutions in $\mathcal{L}_C$ on the sender’s commitment store or attempt to persuade each other about their beliefs — thereby facilitating the adaptation of the agents’ trust models.

If we look at the argumentation tree of Figure 1, then we see that, first of all, the recommendation-seeking agent must communicate its request, using request_recommendation($Jim, \gamma$) with $\gamma$ the goal for which it wants to know whether Jim is trustworthy or not.

The recommendation-supplier answers with assert($trust(Jim, 5)$), which adds the atomic sentence $trust(Jim, 5)$ to its commitment store: $C_R = \{trust(Jim, 5)\}$ and $C_Q = \emptyset$. The seeker can now choose a next action, but there is only one action that makes sense: challenge($trust(Jim, 5)$), resulting in the recommendation-supplier answering justify($trust(Jim, 5), \{E_1, E_3, R_1\}$). The sentences it uses as justification are also added to its commitment store, so we have: $C_R = \{E_1, E_2, E_3, R_1\}$ Now the recommendation-seeker really does have a choice in its next move in the dialogue. We define moves and legal moves in the dialogue next. Not all locutions can be uttered at any moment, there are rules to the dialogue. These are defined by the protocol $P$ in terms of the moves allowed.

Definition 3 (Moves and dialogues (adapted from Prakken’s Definition 5 [10])). The set $M$ of moves in a dialogue is defined as $\mathbb{N} \times \{R, Q\} \times \mathcal{L}_C$, where the three elements of a move $m$ are denoted by, respectively:

- id$(m)$, the numerical identifier of the move
- player$(m)$, the agent performing in the move
- speech$(m)$, the speech act performed in the move

The set of dialogues, denoted by $M^{\leq \infty}$, is the set of all sequences $m_1, \ldots$ from $M$, such that each $i^{th}$ element in the sequence has identifier $i$ and for any $i > 1$, player$(m_i) \neq$ player$(m_{i-1})$\(^3\). The set of finite dialogues is denoted by $M^{< \infty}$. For any dialogue $d = m_1, \ldots, m_i$, the sequence $m_1, \ldots, m_i$ is denoted by $d_i$, where $d_0$ denotes the empty dialogue. When $d$ is a dialogue and $m$ a move, then $d; m$ denotes the continuation of $d$ with $m$.

A protocol $P$ on a set of moves $M$ is a set $P \subseteq M^{< \infty}$ satisfying the condition that whenever $d \in P$, so are all initial sequences of $d$. We define a partial function $P_F : M^{< \infty} \rightarrow 2^M$ for personalised trust dialogues, that allows us to derive the protocol $P$. Prakken defines this in the opposite manner: with the protocol defining the function [10]. In practice, however, it is easier to define the function than all possible sequences of legal moves.

\(^3\) Note that this is a specific implementation of the turn-taking function in Prakken’s dialogue system [10].
**Locution** | **Precondition.** *d* is the dialogue so far and *s* the player request
---|---
request_recommendation(*t, γ*) | A recommendation-seeker may only request a recommendation in the first move, *t* must be a target and *γ* a goal. Formally: $d = d_0$, *t* ∈ Agents and *γ* ∈ $\mathcal{L}_{\text{Goal}}$.

assert(*p*) | A recommendation-supplier may only assert a trust evaluation in the second move, and the goal and target of the recommended trust evaluation must be equal to the goal and target for which it was requested. Formally: $d = m_1$, player(*m*$_1$) = $\exists$, *p* ∈ $\mathcal{L}_{\text{Rep}}$, target(*p*) = *t* and goal(*p*) = *γ*, with *γ* and *t* the goal and target in speech(*m*$_1$), and functions goal and target return the goal and target for which an evaluation is made.

justify(*p*, $S$) | A sentence *p* can be justified, if it is in the current player’s commitment store and the other player challenged it in a previous move. Formally: let $d = d_{i-1}; m_i$ and $\exists = \text{player}(m_i)$, then there is a move *m* in *d*, such that player(*m*) = $\exists$ and speech(*m*) = justify(*p*). Furthermore $p \in C_s$, $S \vdash \text{Arg} p$ and $S \not\subseteq C_s$.

challenge(*p*) | A sentence *p* can be challenged, if it is in the other player’s commitment store and the current player has not previously challenged it. Formally: let $d = d_{i-1}; m_i$ and $\exists = \text{player}(m_i)$, then there is no move *m* in *d* such that player(*m*) = *s* and speech(*m*) = challenge(*p*). Furthermore $p \in C_s$.

counter(*π*$_1$, *π*$_2$) | A priority *π*$_1$ can be countered by priority *π*$_2$, if it is in the other player’s commitment store and *π*$_2$ is not yet in the current player’s commitment store. Formally: let $d = d_{i-1}; m_i$ and $\exists = \text{player}(m_i)$, then *π*$_1$ ∈ $C_\exists$, *π*$_2$ ∉ $C_s$ and *π*$_1, \pi_2 \in \mathcal{L}_{PL}$.

argue(*ψ*) | The current player may propose to argue about belief *ψ* if *ψ* is in the other player’s commitment store and the player has not previously proposed to argue about *ψ*. Formally: let $d = d_{i-1}; m_i$ and $\exists = \text{player}(m_i)$, then there is no move *m* in *d* such that speech(*m*) = argue(*ψ*). Furthermore $\psi \in C_\exists$ and $\psi \in \mathcal{L}_{Bel}$.

end | A player may always choose to end the dialogue after the first move. Formally: $d \neq d_0$.

| Table 3. The preconditions, in terms of the dialogue, for the various locations in $\mathcal{L}_C$ | |
Definition 4 (Protocol function for Dialogues for Recommending Trust).

Pr : \( M^{<\infty} \rightarrow 2^M \) defines the set of legal moves in a dialogue, and thus by induction defines the protocol \( P \) of a dialogue. We do this, by first defining the preconditions for each of the possible speech acts. These are listed in Table 3. We define the function \( \text{pre} \) that, given a speech act, a player and a dialogue, returns whether the preconditions are true or false. This allows us to define a function that returns all legal moves, given the dialogue so far:

- \( \text{Pr}(d_0) = \{ (1, Q, \text{recommendation}(t, \gamma, r)) \} \)
- \( \text{Pr}(d; m_i) = \{ (i+1, s, \text{lm}) | s = \text{player}(m_i) \land \text{lm} \in L_C \land \text{pre} (\text{lm}, s, d; m_i) \} \)

If the persuasion dialogue about argumentation is guaranteed to terminate, then the dialogue for recommending trust is guaranteed to terminate. The proof of this is trivial, given that \( L_{\text{Arg}} \) contains a finite number of elements and the protocol guarantees no steps are repeated. It depends, however, on the agents’ choices of the legal moves how fast it reaches a desirable outcome. Such a desirable outcome is furthermore dependent on the agents actually adapting their trust models when necessary. This is not treated in the actual dialogue: if either agent receives a trust priority rule as the justification for a priority, it may choose to add this to its own rule base. This is a choice made outside of the dialogue, and if this happens then the argumentative theories change. This means the logic for the current dialogue no longer represents the agents’ stances, and therefore the agent should choose to end the current dialogue. The seeker should restart with a new request for recommendations. We describe the choices an agent can make in more detail in [4].

4 Discussion and Conclusions

The protocol we propose is specially designed for a dialogue with the aim to personalise a trust recommendation. It therefore includes the locutions to counter a priority with another one, and propose to argue about beliefs that underly a trust evaluation. These two locutions are the main points of difference between the dialogue we propose and a standard information-seeking dialogue, such as presented by Prakken [10]. Moreover, we have taken special care to allow agents to choose what information they disclose.

First, at any point in the dialogue, the agent may choose to end the dialogue. Unfortunately, this also means that we can give no formal guarantees about whether a dialogue ends in success or not; an agent may always choose not to disclose an important piece of information for the adaptation process.

Secondly, the dialogue moves through an argument step-by-step. If a sentence in the agent’s commitment store is challenged, the agent only justifies it with the direct reason for believing the sentence: only one node of the argument tree is unfolded at a time, and the agent may choose to end the dialogue, rather than disclose the information at any point. These three properties of our dialogue make it quite different from Pinyol’s dialogue protocol [8], which is, insofar as we know, the only other dialogue protocol specifically designed for communicating about trust. Pinyol’s dialogue protocol is more reminiscent of one for a
persuasion dialogue, because each move communicates an entire argumentation tree and the only legal move is to counter some part of that tree. The dialogue only ends when an agent can no longer attack the other’s arguments, at which point the recommendation-seeker decides whether or not to accept the trust recommendation or not.

The dialogue protocol presented in this paper thus improves on the state of the art in two ways: first, it extends existing dialogue protocols specifically to allow for the adaptation of trust models and the communication of personalised trust recommendations; and second, it allows agents a more fine-grained control over how much information they communicate than the only other argumentation protocol for trust that we know about.

References

Reasoning about Advisors for Seller Selection in E-Marketplaces via POMDPs

Frans A. Oliehoek¹, Ashwini A. Gokhale², and Jie Zhang³

¹ Maastricht University, Maastricht, The Netherlands
² MIT, Cambridge, MA, USA
³ Nanyang Technological University, Singapore

Abstract. In an e-marketplace populated with a large number of sellers, some of which may be dishonest, the selection of good sellers to do business with is crucial but challenging especially when buyers do not have much experience with these sellers. In this paper we introduce the SALE POMDP, a framework for the seller selection problem that allows the decision maker to reason both about the quality of the sellers, as well as the trustworthiness of the advisors. In particular, the framework allows the agent to ask advisors about the trustworthiness of other advisors while still offering the benefit of optimally trading off information gathering and exploitation of knowledge as afforded by a POMDP based approach. Via this model, we present a preliminary investigation on the benefit of reasoning about trustworthiness of advisors. We also demonstrate how this enables incorporation of trust propagation as an integral part of the decision making process.

Keywords: Seller Selection, E-Marketplace, POMDPs.

1 Introduction

We consider the ‘seller selection’ problem in e-marketplaces, where an agent, the buyer, is assigned with the task of purchasing a particular item and needs to decide from which of the agents that offer the item it should order. In order to make this decision, the buyer maintains a belief over the quality levels of the various sellers. Also it can ask peers about their beliefs in order to improve its estimate of the quality levels. Only when the buyer is sufficiently sure that it has identified a seller with sufficient quality, should it go ahead and order the item (so the problem includes the decision of whether to place an order).

There have been a number of approaches to maintaining Bayesian beliefs over the quality levels of sellers, by integrating the buyer’s own beliefs as well as the beliefs of other buyers (advisors) [12, 13]. These approaches tend to focus only on obtaining an accurate estimate of seller quality, but fail to reason about when it is necessary to query advisors in order to make optimal decisions.

An approach was suggested to perform full Bayesian decision making by casting the seller selection problem as a partially observable Markov decision process (POMDP), named Advisor POMDP [6]. POMDPs provide a generic framework
for optimal decision making for an agent in a stochastic and partially observable environment [5]. The advantages of a POMDP approach are as follows: 1) rather than trying to achieve the most accurate estimate of sellers, the approach tries to select good sellers and does that optimally; reasoning about sellers is a means, not an end, and 2) a POMDP approach explicitly reasons about information gaining actions in partially observable environments, which allows the agent to optimally trade off the cost of obtaining and benefit of more information.

However, the Advisor POMDP framework assumes that all advisors are equally trustworthy. Following other approaches [12, 13], we acknowledge that it is important to model the trustworthiness of advisors; modeling the trustworthiness of advisors has a big impact on the optimal policy. We also believe that by allowing the buyer to query about (other) advisors we can integrate trust propagation into the decision making process, and thus improve the approach.

We introduce a new model called the (S)eller & (A)dvisor se(LE)ction POMDP (SALE POMDP), which implements these ideas by explicitly incorporating a model of the advisors’ trustworthiness in the state description. By asking advisors about both sellers and other advisors, a SALE POMDP-agent can improve its belief and subsequently take an informed decision on whether to place an order and if so from which seller. In this paper, we demonstrate how this belief revision process works, show that taking into account trust of the advisors is important in the seller selection problem and that, under certain circumstances, allowing the agent to ask advisors about other advisors allows it to realize a higher expected utility for its owner.

2 Background

2.1 Reputation Systems

Some sellers in e-marketplaces may be dishonest and not deliver products with the quality levels as they promised or declared. Thus, seller selection in such uncertain environments is important. Reputation systems have been introduced to address this issue and are particularly useful when buyers do not have much direct experience with sellers [4]. Among them, Bayesian approaches [12, 13] have drawn large attention. For example, Teacy et al. [12] proposed the TRAVOS model, which is a trust and reputation model based on the beta probability density function, and integrates a buyer’s own beliefs about sellers as well as the beliefs of advisors. However, these approaches do not provide optimal decision making for the buyer on whether and from which seller to place an order, which is exactly what our approach tries to offer.

Those Bayesian approaches also suggest to model the trustworthiness of advisors as some advisors may lie about their experience with sellers. For example, [10] proposes to learn about the advisors by trying to estimate the properties of sellers and using those to estimate the advisor’s advice given those properties. This is on a somewhat different time-scale than our approach. In particular, in order to learn about an advisor, the agent should first have many interactions
with a seller (to be certain enough about the sellers properties) at that point, the agent can learn what type of advice the advisor gives for such seller properties. This means, however, that in order to learn about a single advisor multiple transactions of both our agent and the advisor with the same seller are required.

In contrast, we hope to be able to learn about advisors by asking other advisors, thereby avoiding the need to engage in costly transactions.

2.2 The Advisor POMDP

Regan et al. [6] introduced the Advisor POMDP, an approach for dealing with the seller section problem based on the POMDP framework. Formally, an Advisor POMDP consists of the following elements:

- There are $I$ advisors that can be queried about the reputation of all $J$ sellers.
- $S$—a set of possible states of the environment. A state $s = \langle q, sat \rangle$, where $q \in [0, 1]^J$ is a vector indicating the quality $q_j$ of each seller and $sat \in \{-1, 0, +1\}$ indicates whether the result of a purchase is satisfactory (+1), unsatisfactory (-1) or whether no purchase took place yet (0).
- $A$—a set of actions. There is one action $ask_i$ for each advisor $i$, and one $buy_j$ action for each seller $j$.
- $T$—a transition function that specifies $Pr(s' | s, a)$, the probability of transferring to a state $s'$ given that action $a$ was taken in state $s$. For $ask_i$ actions, the state does not change. For $buy_j$ action a state $s = \langle q, 0 \rangle$ changes stochastically to $s' = \langle q, -1 \rangle$ or $s = \langle q, +1 \rangle$ with probabilities depending on $q_j$.
- $R$—a reward function specifying $R(s, a, s')$. For ask actions, a small cost is paid independent of the state. For transitions to a satisfied state (i.e., from $sat = 0$ to $sat = +1$) a reward is received, while transitions to an unsatisfied state yield a large penalty. Once the state changed to satisfied or unsatisfied, no further rewards are given.
- $\Omega$—a set of observations $o$. In the advisor POMDP, the advisors respond with a tuple $o = \langle rep_j, cf_j \rangle_{j=1}^J$ that expresses the knowledge of that advisor about all sellers. Here $rep_j$ is the reputation according to the advisor and $cf_j$ is a measure of how certain the advisor is.
- $O$—the observation function that specifies $Pr(o | a, s)$. Since the semantics of the certainty factors are not formalized, there is some freedom in its specification.
- $b^0$—the initial state distribution.
- $h$—the horizon of the problem. That is the number of time steps, or stages, for which we want to plan. We will assume that $h$ is infinite in this paper.

When the agent interacts with the environment, it can maintain a so-called belief $b$, i.e., a probability distribution over states via Bayes’ rule. That is, when $b(s)$ specifies the probability of $s$ (for all $s$), we can derive $b'$ an updated belief after taking some action $a$ and receiving an observation $o$. Assuming discrete sets of states and observations (as we will do in the remainder of the paper), this update can be written as follows:

$$b'(s') = \frac{Pr(s', o | b, a)}{Pr(o | b, a)} = \frac{1}{Pr(o | b, a)} Pr(o | a, s') \sum_s Pr(s' | s, a) b(s).$$ (1)
Here, $Pr(o|b, a)$ is a normalization factor.

These beliefs are the basis for decision making: a policy $\pi$ maps beliefs to actions $\pi(b) = a$. The goal of solving the POMDP is to find an optimal policy that maximizes the expected discounted cumulative reward, also called value:

$$V(\pi) = \mathbb{E}\left[\sum_{t=0}^{h-1} \gamma^t R(s, a, s') \mid \pi, b^0\right],$$

(2)

with $0 \leq \gamma < 1$ the discount factor.

Finding an optimal policy $\pi^*$ is intractable in general (PSPACE complete [8]), however, in recent years substantial advances have been made in the approximate solution of POMDPs (e.g., [7, 11]).

### 3 Reasoning about which Advisors to Trust

The Advisor POMDP presents a coherent and principled framework to making decisions in the seller selection problem. However, there are some limitations to this model, as we now discuss.

A severe limitation is that the Advisor POMDP puts equal trust in all the advisors. In a real system it is absolutely not a priori clear that all advisors can be trusted and we hypothesize that this may have a big impact on how one should act (i.e., what the optimal policy is). In fact, there is a large field of research on trust propagation that deals with the question of how one should adapt the trust in peers [3, 2]. A disadvantage of current approaches, however, is that they deal with the problem of most accurately estimating the trust levels, rather than integrating this type of reasoning with the decision process of selecting a seller. As a result, it is not clear how one would actually optimally apply such approaches in the context of seller selection. Here we try to overcome this problem by presenting a new model that incorporates these ideas from trust propagation within a POMDP formulation.

Also, in the Advisor POMDP, each advisor gives its ratings about all the sellers. However, instead of estimating the quality of all sellers, the only goal should be to select the seller with high quality. As such, the observation in the advisor POMDP may contain a lot of unnecessary information, leading to unnecessary communication. Therefore we will consider an approach in which our agent has to indicate about which seller (or other advisor) it wants to ask.

These ideas lead us to the formulation of a new model called the *(S)eller *(A)dvisor se(LE)ction POMDP *(SALE POMDP)*, which we will formally introduce in the next section. Section 3.2 will present an example instantiation of the framework that we use in our experimental evaluation.

---

4 By using a different observation function it would be possible to have observations from different advisors result in different beliefs, thereby modeling different levels of trust. These levels, however, would be assumed known to the agent.
3.1 The SALE POMDP Model

Like the Advisor POMDP, the SALE POMDP is a sub-class of POMDP problems. On the one hand the SALE POMDP is more complex than the Advisor POMDP: we assume that the advisors also have a quality, or trustworthiness, and that this is part of the state space. Moreover we introduce extra actions as we allow the agent to ask about the quality of other advisors. On the other hand, we make the simplifying assumption of having discrete sets of quality levels, which allows us to use standard POMDP solvers.

Since the SALE POMDP is a POMDP, it can be described in terms of states, actions, observations and rewards.

**States.** Like in the Advisor POMDP, a state contains the quality levels of all sellers, however, it also contains the quality, or trustworthiness, of each advisor. Let \( Q \) be the discrete set of seller quality levels and \( U \) be the set of advisor quality levels. Then, a state is a tuple \( s = \langle q, u, sat \rangle \), where \( q \in Q \) is a vector indicating the quality of each seller, \( u \in U \) a vector indicating the quality of each advisor, and \( sat \in \{-1, 0, +1\} \) as before. We also write \( q_j \) for the \( j \)-th element of \( q \) and \( u_i \) for the \( i \)-th element of \( u \). After a buy action is taken, the decision process ends. This is modeled using sets of terminal states. That is, a terminal state is a state where \( sat = +1 \) or \( sat = -1 \). We will think of these sets of states as single states called satisfied and unsatisfied.

**Actions.** The model knows the following types of actions:
- \( seller \text{-} query_{ij} \) — ask advisor \( i \) about seller \( j \),
- \( advisor \text{-} query_{ii'} \) — ask advisor \( i \) about advisor \( i' \),
- \( buy_j \) — buy from seller \( j \).
- \( do \text{-} not \text{ buy} \) — decide not to buy from any seller.

**Transitions.** As in the Advisor POMDP, we assume that when taking a query action, the state does not change:

\[
\forall i, j \quad \Pr(s'|s, seller \text{-} query_{ij}) = \delta_{ss'}, \\
\forall i, i' \quad \Pr(s'|s, advisor \text{-} query_{ii'}) = \delta_{ss'},
\]

where \( \delta_{ss'} \) is the Kronecker delta that is 1 if and only if \( s = s' \).

When taking a \( buy_j \) action, the state will always transition to a terminal state. The transition probabilities to terminal states give a definition of the quality levels. In general, chances of transitioning to ‘satisfied’ should be higher when buying from higher quality sellers \( j \).

Together, the specifications of these transitions imply the assumption that quality and trust-levels are stationary for the duration of the decision process.

**Rewards.** The SALE POMDP specifies the following rewards: A small cost associated with ask actions \( R(s, seller \text{-} query_{ij}) = R(s, advisor \text{-} query_{ii'}) = R_{ask} \), a reward associated with a good purchase \( R(s, buy_j, s' = \langle q, u, sat = +1 \rangle) = R_{sat} \), and a penalty associated with dissatisfaction \( R(s, buy_j, s' = \langle q, u, sat = -1 \rangle) = R_{unsat} \). There is a penalty associated with taking the \( do \text{-} not \text{ buy} \) action when in
fact there is a seller of high enough quality (we use $-R_{\text{sat}}$), otherwise the reward for this action is 0.

**Observations.** When a query action is performed the agent will receive an observation from the set of discriminated quality levels. That is, after a seller query action, the agent receives an observation $o \in Q$ corresponding to the quality of seller $j$, while after an advisor query the agent will get an observation $o \in \mathcal{U}$ corresponding to the quality of the advisor $i'$. When the agent transitions to a terminal state, it receives the observation ‘ended’. As such $\mathcal{O} = Q \cup \mathcal{U} \cup \{\text{ended}\}$.

As in the Advisor POMDP, there is no a priori correct way to specify the observation probabilities. In fact, the probabilities picked for the observation function define the meaning of different trust levels. In general, the idea is that trustworthy advisors will give more accurate and consistent answers than untrustworthy ones.

**Initial State Distribution.** The initial state distribution is dependent on the subjective beliefs of the agent (or its owner) when the need for purchasing an item arises. In the case that nothing is known, it makes sense to start with a uniform belief over the quality levels, but a different initial belief could have resulted from previous interactions.

That is, once the buy action is taken, the resulting belief can be used as the basis for an initial belief for a new seller selection instantiation. There are two sources of previous experience: 1) Previous seller selection tasks: the modified belief state resulting from advice in a previous problem can be retained, and 2) Actual experiences with sellers: even though in the decision making task we model a transition to a terminal state with a deterministic ended observation, the actual order will result in the owner of the agent being satisfied or not and this information can be used to update the final belief of the agent’s previous seller selection task giving a new initial belief for a new task.\footnote{In fact this can be an important mechanism to deal with advisors that are consistent but deceptive and settings in which the majority of advisors is untrustworthy.}

### 3.2 Example

Suppose that there are $J = 2$ sellers for the item in concern, each of which can have $|Q| = 2$ quality levels. In this example we use $Q = \{L, H\}$ for low and high quality. Then we have $2^J = 4$ possible ‘quality states’ $q$:

$$q \in \{(L, L), (L, H), (H, L), (H, H)\}. \quad (5)$$

Also suppose that there are $I = 3$ advisors, each of which is T(rustworthy) or U(ntrustworthy). That is $\mathcal{U} = \{T, U\}$. This leads to $2^I = 8$ ‘trust states’ $u$:

$$u \in \{(T, T, T), (T, T, U), \ldots, (U, U, U)\}. \quad (6)$$

As such, an example of a fully specified state is $s = (\langle L, H \rangle, (T, U, U), 0)$.

The transition function for the query actions is specified as explained: the underlying state does not change. There is some freedom in specifying the transition...
Table 1: Observation probabilities.

<table>
<thead>
<tr>
<th></th>
<th>good</th>
<th>bad</th>
<th></th>
<th>good</th>
<th>bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>T H</td>
<td>0.9</td>
<td>0.1</td>
<td>T T</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>T L</td>
<td>0.1</td>
<td>0.9</td>
<td>T U</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>U H</td>
<td>0.5</td>
<td>0.5</td>
<td>U T</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>U L</td>
<td>0.5</td>
<td>0.5</td>
<td>U U</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

(a) $Pr(o|seller\_query_{ij}, q_j, u_i)$ for the possible observations 'good' and 'bad'.

(b) $Pr(o|advisor\_query_{i'j}, u_i, u_{i'})$ for the possible observations 'good' and 'bad'.

In our example, we label the two possible observations 'good' (i.e., the advisor says that the seller/other advisor is high quality) and 'bad' (the seller/other advisor is said to be low quality). As mentioned above, the observation probabilities when transferring to a terminal state are fixed: the agent will observe ended with probability 1. The observation probabilities for the $seller\_query_{ij}$ action (ask advisor $i$ about seller $j$) should be such that asking a trustworthy advisor $i$ gives more accurate observations. One possible way to specify these probabilities is shown in Table 1a. Similarly, Table 1b shows example observation probabilities for the $advisor\_query_{i'j}$ action.

4 Experiments

In this section we report upon a first empirical investigation of the SALE POMDP model. In particular, we demonstrate how the belief update (1) leads to correlation of particular states which forms the basis of improved decision making. We also show that in our example setting it is important to explicitly take into account the advisor’s trustworthiness and that asking advisors about other advisors can be beneficial in certain settings.

In order to perform the empirical evaluation we utilize SARSOP [7], a state-of-the-art POMDP solver, which reads in problems in a standardized POMDP description format. SARSOP does not exploit the factored structure of our problem, therefore, in specifying our models, we substituted all the terminals by two separate states satisfied, unsatisfied, reducing the number of states. Furthermore, the models we used specified two quality and trust levels as in the example of Section 3.2. Also, unless noted otherwise, the transition and observation models used the same parameters as described in that section. For the rewards, we used $R_{ask} = -1, R_{sat} = 50, R_{unsat} = -100$. Also, we penalized taking the $do\_not\_buy$ action from states where there was a high quality seller with $-50$. 

73
4.1 Illustration of Belief Update: Correlation between States

Here we provide some intuition behind the SALE POMDP model by illustrating the process of belief updating. The basic idea is that the belief updates should correlate the state factors in meaningful ways. For instance, observing \textit{good} after \textit{seller\_query}_{ij} should give more weights to states where the seller is high quality \( q_j = H \) and the advisor is trustworthy \( u_i = T \), and less weights to states where the seller is low quality \( q_j = L \) and the advisor is trustworthy \( u_i = U \). This is clearly demonstrated in a number of transitions in Figure 1b, which shows the policy found for a one seller (\( J = 1 \)) one advisor (\( I = 1 \)) SALE POMDP. Similarly, observing \( T_j \) after \textit{advisor\_query}_{ii^{'}} should put more weight on states where \( u_{i^{'}} = T \) and \( u_i = T \), and decrease weight on states where \( u_{i^{'}} = U \) and \( u_i = T \).

4.2 The Impact of Taking into Account Trust

Here we compare (a simplified version of) the Advisor POMDP with one seller (\( J = 1 \)) and one advisor (\( I = 1 \)) to the SALE POMDP model. For both models we use the same discretization of quality levels (\( Q = \{L, H\} \)), so the only difference is that the SALE POMDP includes the trustworthiness of the advisor as a state variable and that the observations are dependent on this factor. The observation model is as shown in Table 1.

Figure 1 shows the policies found for the two models. It clearly shows that the policies are qualitatively different. In particular, while in the Advisor POMDP it is possible to return to the initial belief after observing an equal number of \textquote{good} and \textquote{bad} observations. In contrast, in the SALE POMDP this leads to a belief where the advisor is thought to be untrustworthy. As such, the agent is able to reason about the trustworthiness of the advisor by repeated interactions.

Since, the Advisor POMDP corresponds to the setting in which there is a single trustworthy advisor, it achieves higher value (mean value of 1000 evaluations is 5.36) than the SALE POMDP (–8.56). However, the policy found for the Advisor POMDP with only one untrustworthy advisor (i.e., with \textquote{advice accuracy} 0.5) is much lower (–19.88). Interestingly, the mean of 5.36 and –19.88 (–7.26) corresponds to the setting where when an advisor type is chosen with 50\% probability and then revealed to the agent. We see that the SALE POMDP achieves value fairly close to this \textquote{oracle} upper bound.

4.3 Multiple Advisors: Trust Propagation

We also hypothesize that allowing the agent to query advisors about other advisors, thereby integrating a form of trust propagation in the seller selection decision procedure, can allow for further improvements. In order to test this hypothesis, we consider the SALE POMDP framework with three advisors and compare it to a baseline model: the same SALE POMDP model but without the \textit{advisor\_query}_{ii^{'}} actions, which we will call the NoAQ model. The top row of Table 2 lists the results of this comparison. In contrast to our expectation,
Fig. 1: Comparison of policies found. Thick nodes indicate where *buy* and *do not buy* actions are taken. Nodes also show the belief over (non-terminal) states.

we see that the NoAQ model actually performs better. However, since the set of policies for the regular SALE POMDP model is a strict superset of those for the NoAQ model, we know that the former should be able to achieve at least the same value. The fact that this does not happen can be attributed to the additional computational complexity (induced by the additional actions).

The fact that the regular model does not outperform NoAQ also means that the latter is able to sufficiently figure out which advisors are trustworthy using only *seller_query* actions (as also discussed in Sect. 4.2). Therefore we form a new hypothesis that asking about advisors is beneficial when *seller_query* actions do not provide much information about the trustworthiness of an advisor. This is confirmed by the other test results shown in the table that show what happens if the accuracy with which trustworthy advisors report about sellers (i.e., the ‘0.9’ from Table 1a) diminishes.

The mentioned additional computational complexity of the SALE POMDP model is further demonstrated in the bottom part of the table. It shows that allowing for additional solution time over 100s is improving the quality of the policy further, while for NoAQ the further improvement is marginal.
5 Discussion & Future Work

Here we discuss the limitations of the SALE POMDP model and point to avenues for future work.

5.1 Overcoming Restricting Modeling Assumptions

A restrictive assumption is that we assume independent responses from the same advisor. That is a response of an advisor does not depend on an earlier response from that same advisor. While this is a common assumption (e.g. [10]), it might not be the most realistic. This restriction can be overcome by modeling a previous response as part of the state. The state factor for an advisor \(i\) could for instance have value \(\langle\text{trustworthy, not queried yet}\rangle\) or \(\langle\text{untrustworthy, seller-3-high-quality}\rangle\).

The current model can also be a good model for groups of advisors, when we ask a different person within that particular group (e.g., IP-range).

Another issue is that we are currently assuming a simplistic form of untrustworthy: basically (more) random. However, in real-life, untrustworthy advisors may give very biased answers. In such cases it is much more difficult to identify untrustworthy advisors, but this would also be a problem for a human decision maker. Still, the result of the final belief update (transferring to an unsatisfied state), will correct for the wrong belief that that advisor was trustworthy. Moreover, other approaches to dealing with deceptive advisors based on Bayesian updating (such as [10]) can be neatly integrated in our approach.

5.2 Scaling Up

Solving POMDPs is intractable in general, but in recent years, huge advances have been made in approximate solutions: good policies have been found for

Table 2: Results for the SALE POMDP with multiple advisors. Shown are mean values from 1000 evaluations, together with 95% confidence bounds.

<table>
<thead>
<tr>
<th></th>
<th>regular</th>
<th></th>
<th></th>
<th>NoAQ</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>LB</td>
<td>UB</td>
<td>mean</td>
<td>LB</td>
<td>UB</td>
</tr>
<tr>
<td>Varying accuracy of advisors w.r.t. seller quality (1000s) accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>-3.24</td>
<td>-3.67</td>
<td>-2.81</td>
<td>-2.87</td>
<td>-3.30</td>
<td>-2.44</td>
</tr>
<tr>
<td>0.8</td>
<td>-5.31</td>
<td>-5.74</td>
<td>-4.87</td>
<td>-6.39</td>
<td>-6.79</td>
<td>-5.99</td>
</tr>
<tr>
<td>0.7</td>
<td>-8.08</td>
<td>-8.43</td>
<td>-7.73</td>
<td>-9.77</td>
<td>-10.10</td>
<td>-9.44</td>
</tr>
<tr>
<td>0.6</td>
<td>-13.20</td>
<td>-13.50</td>
<td>-12.90</td>
<td>-14.82</td>
<td>-15.09</td>
<td>-14.56</td>
</tr>
<tr>
<td>Impact of solution time (0.7 accuracy) time(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>-8.08</td>
<td>-8.43</td>
<td>-7.73</td>
<td>-9.77</td>
<td>-10.10</td>
<td>-9.44</td>
</tr>
</tbody>
</table>
problems with thousands of states. Still, our empirical evaluation indicates that computational costs are a limiting factor in the use of our framework. However, there are a number of different forms of structure in the proposed POMDP formulation that can potentially be exploited for faster solutions. For instance, the SALE POMDP really is a factored POMDP, so we may try and use solvers that exploit this property such as symbolic Perseus [9]. Moreover, there is other special structure that we may be able to exploit. For instance, we do not care about which seller we end up ordering from, but only about whether that seller has sufficient quality. This means that there are symmetries between sets of states (e.g., a \( \langle H, L, T \rangle \) has the same value as \( \langle L, H, T \rangle \)) which may be exploited [1]. Moreover, there may be little difference in value between (beliefs assigning high probability to) states that have one high-quality seller and states that have multiple high quality sellers. In future work we hope to exploit these types of structure for improved computational efficiency.

5.3 Learning Accurate Models

For our proof-of-concept experiments, we specified all the parameters in an ad-hoc fashion. However, it is important to know that all those numbers can be estimated in a sensible way. Moreover, for a decentralized peer-to-peer type of system to work, it is required that each peer (each agent) can adapt its model by learning because 1) it needs to be able to adapt to changes over time, and 2) it needs to adapt to the preferences of its owner [10].

Learning POMDP models is a very difficult problem, but there are some special properties that could be exploited. First, the observation model might not need to be learned since it in fact encodes the definition of our different trust-levels. Second, the state of the world in this problem does not change (or perhaps only very slowly), which may lead to easier learning. Finally, we would like to point out that even if the model is not completely accurate, it is still possible to get out very high quality policies: The general form of “buy when certain enough about one seller” does not change. For instance, by setting the reward function conservatively, we know that the agent might perform too many queries in case that the model was right, but it also builds in a robustness against the uncertainty in model parameters.

6 Conclusion

In this paper, we proposed a novel (S)eller & (A)dvisor se(LE)ction POMDP model (SALE POMDP) to address the problem of seller selection in e-commerce settings. Our model provides a principled approach for buyers to optimally select sellers trading off the cost and benefit of seeking more information from advisors. We provided examples to demonstrate step by step how our model works, and experiments to demonstrate its effectiveness. Encouraged by this promising first

\[6 \text{An interesting other question is how we may optimize these definitions.}\]
step, we also discussed several important next steps to improve our model and increase its applicability.

Acknowledgments

This work was supported, in part by AFOSR MURI project #FA9550-09-1-0538, in part by NWO CATCH project #640.005.003, and in part by Singapore Ministry of Education under a grant to the Singapore-MIT International Design Center.

References


Abstract. Information provided by a source should be assessed by an intelligent agent on the basis of several criteria: most notably, its content and the trust one has in its source. In turn, the observed quality of information should feed back on the assessment of its source, and such feedback should intelligently distribute among different features of the source—e.g., competence and sincerity. We propose a formal framework in which trust is not treated as a monolithic and static concept. We regard trust as a multi-dimensional concept relativized to the sincerity of the source and its competence with respect to specific domains: both these aspects influence the assessment of the information, and also determine a feedback on the trustworthiness degree of its source. We provide a framework to describe the combined effects of competence and sincerity on the perceived quality of information. We focus on the feedback dynamics from information quality to source evaluation, highlighting the role that uncertainty reduction, and social comparison play in determining the amount and the distribution of feedback.

Category: I.2.11, Distributed Artificial Intelligence, Intelligent agents

Terms: Theory

Keywords: Knowledge representation, Cognition, Reasoning

1 Introduction

In real life, the trust assigned to a message depends crucially, albeit not solely, on the perceived trustworthiness of its source [7, 8]. In turn, whether or not the message turns out to be reliable has important repercussions on the source trustworthiness. This is also true with respect to the exchange of arguments in social interaction. When people argue with each other, trying to get their arguments
accepted to foster their own goals, they also evaluate the arguments proposed by other parties in the discussion. This evaluation considers not only the possible conflicts with other arguments, but also the trustworthiness degree of the information source proposing the arguments. In argumentation theory [15], the arguments are considered to be accepted or not depending on the attacks against them. In this kind of frameworks, neither the information sources proposing the arguments nor their trustworthiness degree are considered. In recent years, the area has seen a number of proposals [14, 16, 11, 13, 18, 3] to introduce the trust component in the evaluation process of the arguments. The common drawback of these approaches is that they do not return the intrinsic complexity of the trust notion, as highlighted instead by socio-cognitive models.

In this paper, we adopt the socio-cognitive model of trust proposed by Castelfranchi and Falcone [2], and we elaborate its computational counterpart, although with some expressivity limitations due to the overall complexity of the model. We are interested in investigating how that socio-cognitive model of trust can be extended to a model based on argumentation. This breaks down into the following sub-questions:

- How to distinguish different dimensions of trust, e.g., sincerity and competence, and model their respective contribution?
- How to model the trust feedback from arguments to information sources?

We address these questions starting from the argumentation-based model for belief revision recently proposed by da Costa Pereira et al. [3], which originally lacked a representation of a cognitive model of trust and its inherent dynamics. We extend that model along the following lines.

First, trust is not a monolithic concept. We relativize the notion of trust to the dimensions of sincerity and competence in various domains. For instance, a reliable motor mechanic will be considered competent in the cars domain, but not necessarily so when suggesting the best restaurant to eat pizza; conversely, a pizza maker is typically assumed to be trustworthy on the latter domain but not on the former. The trust in the information source is not absolute, but it is relative to an estimate of sincerity and competence in the relevant domain. Sincerity and competence thus combine to produce an aggregated degree of trust.

Second, trust is not a static concept. There is a bidirectional link between the source and its information items. This means that an argument is more or less believable on the basis of the source’s trustworthiness, but this leads to a feedback such that the invalidation of the argument, due to attacks by other trustworthy arguments, feeds back on the source’s credibility. The sign of the feedback depends on how much the “quality” of the message surprises the agent w.r.t. its prior assessment of the source trustworthiness.

The paper is organized as follows: Section 2 highlights the main differences of our approach with related work. Section 3 provides the basic concepts of the model proposed by Pereira et al. [3]. Sections 4 and 5 introduce the multidimensional trust model and specify the feedback mechanism. Conclusions end the paper.
2 Related Work

The importance of relating trust and argumentation has been underlined by Dix et al. [4], who present trust as a major issue concerning the research challenges for argumentation. Also Parsons et al. [12] present the provenance of trust as one of the mechanisms to be investigated in argumentation. They claim that a problem, particularly of abstract approaches such as Dung [6], is that they cannot express the provenance of trust, and the fact that $b$ is attacked because $b$ is proposed by agent $s$ who is not trustworthy. Starting from this observation, we propose a model of argumentation where the arguments are related to the sources and their degree of acceptability is computed on the basis of the trustworthiness degree of the sources. Furthermore, our approach goes beyond this observation by providing a feedback such that the final quality of the arguments influences the source evaluation as well.

Most studies in this domain used argumentation to model trust dynamics and/or reasoning about trust (e.g., [14, 11, 16, 18]), which is a worthy but different enterprise from the one pursued here. Closer in spirit to the present paper is the work by Parsons et al. [17, 13], who present a framework to introduce the sources in argumentation and to express the degrees of trust. They define trust-extended argumentation graphs in which each premise, inference rule, and conclusion is associated to the trustworthiness degree of the source proposing it. Thus, given two arguments rebutting each other, the argument whose conclusion has a higher trust value is accepted. The difference is that in such a framework the trust values associated to the arguments do not change and the arguments are accepted with the same degree even if they are attacked by more trusted arguments. Again, the feedback towards the source as well as the distinction between competence and sincerity is not considered.

A huge amount of research has been conducted on trust, and some of these works are described below, even if in this paper we limit our attention to the cognitive trust model of Castelfranchi and Falcone [2]. An approach to model trust using modal logic is proposed by Lorini and Demolombe [10], who present a concept of trust that integrates the trusters goal, the trustees action ensuring the achievement of the trusters goal, and the trustees ability and intention to do this action—taking again inspiration from [2]. Another proposal is presented by Liau [9], in which the influence of trust on the assimilation of information into the source’s mind is considered. The idea is that “if agent $i$ believes that agent $j$ has told him the truth on $p$, and he trusts the judgement of $j$ on $p$, then he will also believe $p$”. Wang and Singh [20], instead, understand trust in terms of belief and certainty: $A$’s trust in $B$ is reflected in the strength of $A$’s belief that $B$ is trustworthy. They formulate certainty in terms of evidence based on a statistical measure defined over a probability distribution of positive outcomes. Both Liau [9] and Wang and Singh [20] capture intuitions that play a role also in our approach, but they vastly oversimplify the nature and dynamics of trust, as opposed to the socio-cognitive model discussed in Castelfranchi and Falcone [2].
3 Background

A classical propositional language may be used to represent information for manipulation by a cognitive agent.

Definition 1. (Language) Let $\operatorname{Prop}$ be a finite set of atomic propositions and let $\mathcal{L}$ be the propositional language such that $\operatorname{Prop} \cup \{\top, \bot\} \subseteq \mathcal{L}$, and, $\forall \phi, \psi \in \mathcal{L}$, $\neg \phi \in \mathcal{L}$, $\phi \land \psi \in \mathcal{L}$, $\phi \lor \psi \in \mathcal{L}$.

As usual, one may define additional logical connectives and consider them as useful shorthands for combinations of connectives of $\mathcal{L}$, e.g., $\phi \supset \psi \equiv \neg \phi \lor \psi$. We denote by $\Omega = \{0, 1\}^{\operatorname{Prop}}$ the set of all interpretations on $\operatorname{Prop}$. An interpretation $I \in \Omega$ is a function $I : \operatorname{Prop} \rightarrow \{0, 1\}$ assigning a truth value $p_I$ to every atomic proposition $p \in \operatorname{Prop}$ and, by extension, a truth value $\phi_I$ to all formulas $\phi \in \mathcal{L}$.

We denote by $[\phi]$ the set of all models of $\phi$, $[\phi] = \{I : I \models \phi\}$.

A Dung’s abstract argumentation framework [6] ($AF$) is a pair $\langle A, \rightarrow \rangle$ where $A$ is a set of elements called arguments and $\rightarrow \subseteq A \times A$ is a binary relation called attack. Dung defines a number of acceptability semantics [6] to assess which are the sets of accepted arguments. We can give arguments a structure, and the attack relation is defined in terms of such a structure of the arguments: an argument is a pair $A = \langle \Phi, \phi \rangle$, with $\phi \in \mathcal{L}$ and $\Phi \subseteq \mathcal{L}$, such that (i) $\Phi \not\models \bot$, (ii) $\Phi \models \phi$, (iii) $\Phi$ is minimal w.r.t. set inclusion. We call $\phi$ the conclusion and $\Phi$ the support of the argument. Given $A \in A$, $\operatorname{src}(A)$ is the set of the sources of $A$.

In a recent paper, da Costa Pereira et al. [3] propose a framework where argumentation theory is used in belief revision. In this framework, the arguments are weighted on the basis of the trustworthiness degree of the agents proposing them. The acceptability of the arguments is then computed by a labelling algorithm which assigns the arguments a fuzzy value, differently from Dung-like frameworks where arguments are either accepted or rejected. We select this work as the basis of our trust model because it provides (i) an explicit link between the trustworthiness degree of the sources and of the arguments, (ii) a mechanism such that the initial value assigned to the arguments changes due to the attacks against them, (iii) the beliefs of the agents are also involved, not only their arguments.

Given an $AF$ and the trust degree $\tau_s$ of each source $s$, the labeling algorithm [3] computes a fuzzy extension as a fuzzy set of accepted arguments, whose membership $\alpha$ assigns to each argument $A$ a degree of acceptability $\alpha(A)$ such that $\alpha(A) = 0$ means the argument is outright unacceptable, $\alpha(A) = 1$ means the argument is fully acceptable, and all cases inbetween are provided for. Then the labeling $\alpha$ is used to determine the agent’s beliefs, by constructing a possibility distribution from which the degree of belief of an arbitrary formula may be calculated.

A possibility distribution may be defined as the membership function of a fuzzy set that describes the more or less possible and mutually exclusive values of one (or more) variable(s) [21]. Indeed, if $F$ designates the fuzzy set of possible values of a variable $X$, $\pi_X = \mu_F$ is called the possibility distribution associated
The identity $\mu_F(v) = \pi_X(v)$ means that the membership degree of $v$ to $F$ is equal to the possibility degree of $X$ being equal to $v$ when all we know about $X$ is that its value is in $F$. A possibility distribution for which there exists a completely possible value ($\exists v_0 : \pi(v_0) = 1$) is said to be normalized.

**Definition 2. (Possibility and Necessity Measures)** A possibility distribution $\pi$ induces a possibility measure and its dual necessity measure, denoted by $\Pi$ and $N$ respectively. Both measures apply to a crisp set $A$ and are defined as follows:

$$
\Pi(A) = \sup_{s \in A} \pi(s); \\
N(A) = 1 - \Pi(\bar{A}) = \inf_{s \in A} \{1 - \pi(s)\}.
$$

(1) In words, the possibility measure of $A$ corresponds to the greatest of the possibilities associated to its elements; conversely, the necessity measure of $A$ is equivalent to the impossibility of its complement $\bar{A}$.

The beliefs of an agent are thus completely described by a normalized possibility distribution $\pi : \Omega \rightarrow [0, 1]$, which represents a plausibility order of possible states of affairs: $\pi(I)$ is the possibility degree of interpretation $I$.

Given $A = \langle \Phi, \phi \rangle$, let $\text{con}(A)$ denote the conclusion of $A$, i.e., $\text{con}(\langle \Phi, \phi \rangle) = \phi$. The possibility distribution $\pi$ induced by a fuzzy labeling $\alpha$ is constructed by letting, for all interpretation $I$,

$$
\pi(I) = \min\{1, 1 + \max_{A : I \models \text{con}(A)} \alpha(A) - \max_{B : I \not\models \text{con}(B)} \alpha(B)\},
$$

where the first maximum accounts for the most convincing argument compatible with $I$, and the second maximum accounts for the most convincing argument against $I$. A world is possible to an extent proportional to the difference between the most convincing argument supporting it and the most convincing argument against it. The world is considered completely possible if such difference is positive or null, but it is considered less and less possible (or plausible) as such difference grows more and more negative.

The degree to which a given arbitrary formula $\phi \in \mathcal{L}$ is believed is calculated from the possibility distribution induced by the fuzzy argumentation framework as $B(\phi) = N(\langle \phi \rangle) = 1 - \max_{I \not\models \phi} \{\pi(I)\}$, where $B$ may be regarded, at the same time, as a fuzzy modal epistemic operator or as a fuzzy subset of $\mathcal{L}$. Notice that $B(\phi)$ can be computed for any formula $\phi$, not just for formulas that are the conclusion of some argument. E.g., if $A$ is an argument whose conclusion is $p$ and $B$ is an argument whose conclusion is $p \supset q$, and $\alpha(A) = \alpha(B) = 1$, then not only $B(p) = B(p \supset q) = 1$, but also $B(q) = 1$, $B(p \land q) = 1$, etc.

Consequences of the properties of possibility and necessity measures are that $B(\phi) > 0 \Rightarrow B(\neg \phi) = 0$, which means that if the agent somehow believes $\phi$ then it cannot believe $\neg \phi$ at all, and

$$
B(\top) = 1, \\
B(\bot) = 0, \\
B(\phi \land \psi) = \min\{B(\phi), B(\psi)\}, \\
B(\phi \lor \psi) \geq \max\{B(\phi), B(\psi)\}.
$$

83
4 Trust model

In this section, we take a look into the multidimensional nature of trust. For the sake of brevity, we simplify Castelfranchi and Falcone’s model [2], and focus only on two broad categories of relevant features in the source: competence (to what extent the source is deemed able to deliver a correct argument) and sincerity (to what extent the source is considered willing to provide a correct argument), both of which contribute to determine the source’s overall trustworthiness. Importantly, evaluations of competence and sincerity are allowed to change across different domains. For instance, I might think that the colleague who is competing with me for a promotion is likely to be insincere in giving me tips on how to improve my career, and yet there is no reason to doubt his sincerity when he suggests me what movie to watch tonight.

Here, we consider competence and sincerity as two possible dimensions for measuring the trustworthiness of a source. Such dimensions will be represented as graded beliefs, that is, an agent believes, to a given degree, that a source is competent (sincere) with respect to a domain. In order to represent the fact that the agent’s beliefs can be incomplete (an agent may not have beliefs on everything), it is important to make it clear that one may either “believe \( p \)”, “believe \( \neg p \)”, or believe none of them, due to ignorance [5]. In such a case, it must be possible to make the fact that \( B(p) = 0 \) and \( B(\neg p) = 0 \) explicit. This is the reason why the competence and sincerity will be represented by a bipolar pair of beliefs, in competence (sincerity) and in the negation thereof.

The idea is that each source is assigned by the agent a belief of competence and another belief of sincerity, both represented by bipolar values ranging between 0 and 1. These beliefs combine to determine the degree of trustworthiness. What is crucial is that the basic properties of sincerity and competence are kept separate, since the feedback will be designed to affect differently each of them, and only as a consequence will also impact on the source’s aggregate trustworthiness.

4.1 Modeling Competence

Here, the idea is to define a number of competence domains, with respect to which (i) the competence of a source \( s \) is evaluated, yielding the trust an agent has in \( s \) when it offers arguments relevant to any domain—the trust in the competence of \( s \) is a vector \( c(s) \) whose elements may be formally regarded as bipolar degrees of belief \( (c_d^+, c_d^-) \), where \( c_d^+ \) is the degree to which the agent believes the source is competent about domain \( d \), and \( c_d^- \) is the degree to which the agent believes the source is not competent about \( d \), and (ii) the positive and negative feedback reflected on the information source is also domain-specific, e.g., if the doctor tells the agent to go to a restaurant that turns out to be bad,

\[ c_d^+ \text{ and } c_d^- \text{ obey the property, typical of necessities and beliefs, that } c_d^+ > 0 \Rightarrow c_d^- = 0 \quad \text{and, vice versa, } c_d^- > 0 \Rightarrow c_d^+ = 0. \]
this reduces the agent’s confidence in her gastronomic taste, not in her medical skills.

We propose to associate the arguments, by way of the formula in their conclusion, to the competence domains depending on the topic they are talking about. The atomic propositions of the language are mapped to the competence domains with a certain degree, i.e., the degree to which the atomic proposition belongs in the specific domain. The degree to which a propositional formula (argument) belongs in a specific domain might be given by the maximum degree to which the atomic propositions on which the truth value of the formula depends belong to the competence domain.

**Definition 3.** Given $D$ the set of competence domains, and $\text{Prop}$ the set of atomic propositions, the association between atoms and domains, is represented by a fuzzy relation $R \subseteq D \times \text{Prop}$. Given $p \in \text{Prop}$, $d \in D$, the membership degree of the atomic proposition $p$ to domain $d$ is $R(d,p) \in [0,1]$. We now extend $R$ of Definition 3 to $D \times \mathcal{L}$.

**Definition 4.** Let $\phi \in \mathcal{L}$ be a propositional formula and $d \in D$ be a competence domain, then

$$R(d,\phi) = \max_{p \in \text{DS}(\phi)} R(d,p),$$

where $\text{DS}(\phi) = \{ p \in \text{Prop} : \exists I \models p, I' \nmid \models \phi : \phi^I \neq \phi^{I'} \}$ is the determinant set of $\phi$, i.e., the set of all atomic propositions on which the truth of $\phi$ depends.

### 4.2 Modeling Sincerity

The notion of *sincerity* is a property typically attributed to agents with goals and intentions. Talking about the sincerity of an information source is assuming that source is another agent and has intentions, which can be in harmony or in conflict with the goals of the recipient. For example, when I meet my bank’s personal investment advisor to get advice on possible placements of my savings, I should know from the outset that, among her goals, she has the one of maximizing the profits of her employer. Therefore, I may expect that she will be tempted to manipulate my beliefs to lure me into buying financial instruments on which the bank makes a profit.

Furthermore, in general, the sincerity of a source should relate differently to each individual domain: a malicious agent may have an advantage to lie about a domain somehow related with its goals, but has no interest in lying about unrelated domains. Therefore, it would be wrong to regard sincerity as an absolute property of a source. Thus, both competence and sincerity are relativized to individual domains.

Based on the above discussion, we will model the beliefs an agent maintains about the sincerity of its sources as a vector $\sigma(s)$, whose components, associated to individual domains, are bipolar values $\langle \sigma^+_d, \sigma^-_d \rangle$ where $\sigma^+_d \in [0,1]$ represents the degree to which the agent has reasons to believe that source $s$ is sincere about...
whereas $\sigma_d^+ \in [0, 1]$ represents the degree to which the agent has reasons to believe the contrary; $\sigma_d^+ = \sigma_d^-$ represents a status of maximal uncertainty about the sincerity of $s$ about $d$. Since $\sigma_d^+ > 0 \Rightarrow \sigma_d^- = 0$ and $\sigma_d^- > 0 \Rightarrow \sigma_d^+ = 0$.

### 4.3 Aggregating Competence and Sincerity

Competence and sincerity may be aggregated into a single trust value $\tau_d$, used to weight arguments, by recalling that both concepts are formally beliefs, although of a special kind, not induced by the agent’s $AF$, but determined by the internal mechanisms of competence and sincerity evaluation.

In particular, one might argue that a source is trusted to the extent that it is believed to be both competent and sincere. If this is the intended meaning of trust, then, for all domains $d$, the degree to which $s$ is trusted about $d$ is given by

$$
\tau_d(s) = B(\text{competent}(d, s) \land \text{sincere}(d, s))
= \min(B(\text{competent}(d, s)), B(\text{sincere}(d, s))),
$$

where $\text{competent}(d, s)$ means that $s$ is competent about $d$ and $\text{sincere}(d, s)$ means that $s$ is sincere about $d$.

This belief degree may be computed by reconstructing the possibility distribution $\pi$ that induces the beliefs

$$
B(\text{competent}(d, s)) = c_d^+,
B(\neg \text{competent}(d, s)) = c_d^-,
B(\text{sincere}(d, s)) = \sigma_d^+,
B(\neg \text{sincere}(d, s)) = \sigma_d^-.
$$

There are four possible worlds, as far as the competence and sincerity of source $s$ about $d$ goes, namely

$$
I_0 = \text{competent}(d, s) \land \text{sincere}(d, s),
I_1 = \text{competent}(d, s) \land \neg \text{sincere}(d, s),
I_2 = \neg \text{competent}(d, s) \land \text{sincere}(d, s),
I_3 = \neg \text{competent}(d, s) \land \neg \text{sincere}(d, s).
$$

Let us abbreviate $\pi(I_i)$ as $\pi_i$. Since we know that

$$
\max\{\pi_0, \pi_1\} = 1 - c_d^-,
\max\{\pi_2, \pi_3\} = 1 - c_d^+,
\max\{\pi_0, \pi_2\} = 1 - \sigma_d^-,
\max\{\pi_1, \pi_3\} = 1 - \sigma_d^+,
$$
we may solve this system of four equations for the four unknown variables \( \pi_0, \pi_1, \pi_2, \pi_4 \) to get

\[
\begin{align*}
\pi_0 &= 1 - \max\{c_d^-, \sigma_d^-\}, \\
\pi_1 &= 1 - \max\{c_d^+, \sigma_d^+\}, \\
\pi_2 &= 1 - \max\{c_d^+, \sigma_d^-\}, \\
\pi_3 &= 1 - \max\{c_d^-, \sigma_d^+\}.
\end{align*}
\]

Therefore, the single trust value for domain \( d \) is given by the conjunction of how much the agent believes source \( s \) is competent and sincere concerning \( d \)

\[
\tau_d(s) = B(\text{competent}(d, s) \land \text{sincere}(d, s)) = \\
= 1 - II([\neg\text{competent}(d, s) \lor \neg\text{sincere}(d, s)]) = \\
= 1 - \max\{\pi_1, \pi_2, \pi_3\} = \\
= \min\{\max\{c_d^-, \sigma_d^+\}, \max\{c_d^+, \sigma_d^-\}, \max\{c_d^+, \sigma_d^+\}\}.
\]

5 Feedback Dynamics

In this section, we define the feedback dynamics on the sources. As we discussed in the previous sections, the acceptability of the arguments in our model depends on (i) the trustworthiness of the information source proposing the arguments, and (ii) the interactions of the proposed arguments and the other beliefs of the agent. What we want to introduce in this section is the idea that the final acceptability value of the arguments provides a feedback on the trustworthiness degree in the information source from the next interaction.

5.1 Overall Feedback: The role of prediction and surprise

The overall amount and sign (increment or decrement) of the feedback depends on how much the overall quality of the message surprises the agent, with respect to its prior assessment of the source trustworthiness. This captures the principle that information quality should change one’s assessment of its source only when the agent learns something new about the capacity of the source to deliver information of either high or low quality. In other words, there should be a feedback on the source only when the quality of its argument tells me something new about the source’s trustworthiness, revealing my previous opinion to be wrong. Otherwise, the quality of the new argument just confirms my previous assessment of the source, and confirmation, by definition, consolidates a pre-existing judgment, rather than modifying it. This points to the role of prediction in feedback dynamics from arguments to sources, and this prediction is based on the pre-existing degree of trustworthiness of the source of a given argument.

Let \( \tau(s) \) be the current degree of trustworthiness of source \( s \) and \( B(\text{con}(A)) \) the degree of belief, in light of current evidence, in argument \( A \) provided by \( s^2 \).

\[\text{con}(A)\] represent the conclusion of argument \( A \).
Assuming that argument quality \( Q(A) \) is given by \( B(\text{con}(A)) \), then the total amount of feedback \( F_A \) produced by argument \( A \) on source \( s \) is given by

\[
F_A = Q(A) - \tau(s).
\] (8)

The overall feedback \( F_A \) in our framework ranges between \(-1\) (utter disappointment) and \(+1\) (wonderful surprise), and goes to \( 0 \) whenever source \( s \) provides an argument \( A \) whose quality is exactly as expected \((Q(A) = \tau(s))\).

The critical point is that this feedback might affect to a different extent the two components of trustworthiness, to wit, competence and sincerity, depending on the agent’s interpretation of what determined it. Feedback on sources is often specific: the agent does not only register the fact that the source provided information of good (or bad) quality, but it also diagnoses what virtue (or vice) prompted the source to do so. For instance, when I find out that a trusted source provided a poor piece of advice, should I conclude that the source was being deliberately insincere, or should I attribute the incident to a lack of competence? The overall trustworthiness is lowered in either case, but for very different reasons. Conversely, when a poorly estimated source provides surprisingly good information, is this because it stopped deceiving me, or because it became more competent?

5.2 Feedback Distribution: Uncertainty reduction and social comparison

Feedback dynamics face the problem of how to distribute the overall feedback \( F_A \) between competence and sincerity. In this respect, our current formal framework faces an important limitation: there is no way to access semantically the source’s beliefs and goals, so they cannot be invoked to justify the different diagnoses the agent could make of information quality—contrary to what happens in real life, as discussed in [2]. For example, if I attribute to source \( s \) a goal which is currently in conflict with my own, any poor quality information I receive from \( s \) is more likely to be imputed to dishonesty than incompetence. In view of this limitation, the best we can do is to use smart rules-of-thumb that capture interesting (albeit partial) feedback regularities, leaving to future work further refinements on this point.

We propose two independent principles, whose effects sum up in distributing the feedback between competence and sincerity. The first principle focuses on the individual’s previous assessment of a specific source, whereas the second principle compares what the single source is saying to what other sources said, and uses this degree of convergence to shape the feedback distribution. The first principle might be labelled uncertainty reduction: the idea is that a feedback should affect more the dimension for which there is greater uncertainty, that is, such that \( U_c = 1 - \max\{c^+, c^-\} \), for competence, or \( U_\sigma = 1 - \max\{\sigma^+, \sigma^-\} \), for sincerity, is greater. This works well for cases where the agent has formed a clear judgment on one dimension but is in doubt on the other: if I am convinced of your honesty but do not know whether you are competent or not (or vice
versa), then the quality of your argument is likely to be interpreted as evidence for or against your competence (or honesty)—after all, that is what I was not sure about to start with.

Given the total amount of feedback $F_A$ from argument $A$, let $c = \langle c^+, c^- \rangle$ be the prior beliefs in competence, $\sigma = \langle \sigma^+, \sigma^- \rangle$ be the prior beliefs in sincerity, and $F_A(c)$ and $F_A(\sigma)$ the amount of feedback assigned to, respectively, $c$ and $\sigma$. Then the following rule is used to capture the principle of uncertainty reduction:

$$F_A(c) = \frac{F_A U_c}{U_c + U_\sigma}, \quad F_A(\sigma) = \frac{F_A U_\sigma}{U_c + U_\sigma}. \tag{9}$$

This formulation satisfies the following desirable properties: (i) $F_A(c) + F_A(\sigma) = F_A$, (ii) $U_c > U_\sigma \Rightarrow |F_A(c)| > |F_A(\sigma)|$, (iii) $U_c < U_\sigma \Rightarrow |F_A(c)| < |F_A(\sigma)|$, and (iv) $U_c = U_\sigma \Rightarrow F_A(c) = F_A(\sigma) = \frac{1}{2} F_A$.

The second principle might be labeled social comparison: the idea is to compute the degree of convergence of an argument, measured as the number of sources that present that argument or arguments that support it, minus all the sources that are proposing arguments in contrast with it. Then convergence and quality of argument $A$ are compared to determine what dimension of trustworthiness is more affected by the feedback, as follows:

(i) if $A$ is good and convergent, there is a stronger positive feedback on sincerity than on competence (each source vouches for the sincerity of the other, even if they should all turn out to be mistaken);

(ii) if $A$ is good and divergent, there is a stronger positive feedback on competence than on sincerity (an isolated source going against popular wisdom is likely to be on to something, like the biblical vox clamantis in deserto);

(iii) if $A$ is poor and convergent, there is a stronger negative feedback on competence than on sincerity (it is unlikely that everybody is conspiring to fool you, whereas it is more plausible that they are all honestly mistaken);

(iv) if $A$ is poor and divergent, there is a stronger negative feedback on sincerity than on competence (this is the typical case of a malicious or derailed source). 3

More precisely, let $\text{Pro}_A$ be the number of sources claiming argument $A$ or supporting it and $\text{Con}_A$ be the number of sources attacking argument $A$. Then we measure the degree of convergence $k_A$ for $A$ as follows:

$$k_A = \frac{\text{Pro}_A - \text{Con}_A}{\text{Pro}_A + \text{Con}_A}. \tag{10}$$

3 Note that these principles are based on a number of assumptions (most notably, high level of independence and low probability of collusion among sources), and thus are not meant to be universally valid. Rather, they exemplify how simple rules-of-thumb can be identified to regulate feedback distribution, even without any explicit representation of context or agent’s mental states. Testing their validity across various communicative situations (e.g., how much collusion is required to make these heuristics ineffective?) is left as future work.
In the limiting case, when there is no source either supporting or attacking $A$, this indicates that the argument has no external source, hence convergence does not apply. Otherwise, $k_A$ always ranges between 1 (only supporting sources) and $-1$ (only attacking sources), with the 0 value indicating instances where the same number of sources support and attack $A$. Then the following rule is used to capture the principle of social comparison:

- if $k_A > 0$ and $F_A > 0$ (convergent argument producing a positive feedback), then the positive product $k_A F_A(c)$ is added to $F_A(\sigma)$ and subtracted from $F_A(c)$.
- if $k_A < 0$ and $F_A > 0$ (divergent argument producing a positive feedback), then the positive product $k_A F_A(\sigma)$ is added to $F_A(c)$ and subtracted from $F_A(\sigma)$.
- if $k_A > 0$ and $F_A < 0$ (convergent argument producing a negative feedback), then the negative product $k_A F_A(\sigma)$ is added to $F_A(c)$ and subtracted from $F_A(\sigma)$.
- if $k_A < 0$ and $F_A < 0$ (divergent argument producing a negative feedback), then the negative product $k_A F_A(c)$ is added to $F_A(\sigma)$ and subtracted from $F_A(c)$.
- finally, if $k_A = 0$ (neither convergent nor divergent), then social comparison has no effect on feedback distribution.

It is worth noting that the combination of both principles determines a distribution of the feedback between competence and sincerity that respects the following constraints:

$$0 \leq |F_A(c)| \leq |F_A| \quad \text{and} \quad 0 \leq |F_A(\sigma)| \leq |F_A|. \quad (11)$$

This states that distributing the feedback between competence and sincerity neither changes the overall amount of feedback, nor modifies its sign. At most, all the feedback will apply to only one dimension and not at all to the other: this happens, for instance, when the argument is either completely convergent ($k_A = 1$) or divergent ($k_A = -1$), but it never happens that the sum of the feedback on sincerity and the feedback on competence exceeds the overall amount of feedback produced by the argument, nor that the same argument generates a positive feedback on one dimension and a negative feedback for the other one. This is consistent with basic intuitions on how feedback ought to happen.

All in all, the system of rules described in this section provides a relatively simple way to characterize feedback dynamics from information quality to source evaluation, allowing to discriminate multiple dimensions of trustworthiness and capturing some intuitions on how feedback should be distributed among them. We certainly do not claim that these rules-of-thumb are perfect or immune to counter-examples—quite the contrary. However, they strike us as a useful and productive approximation of regularities in feedback dynamics, given current formal limitations in providing a semantic link between context of interaction, agents mental states (beliefs and goals) and argument assessment.
5.3 Feedback Application

We now have all the elements required to define how the feedback produced by argument \( A \) on source \( s \) is applied to the vectors \( c(s) \) and \( \sigma(s) \).

For the sake of generality, let us denote by \( \langle x^+, x^- \rangle \) the bipolar degrees that must be updated, by \( f \) the relevant dimension of feedback, and by \( \langle y^+, y^- \rangle \) the updated degrees. One may regard \( \langle x^+, x^- \rangle \) as formally equivalent to a variable \( x \in [-1, 1] \), defined as

\[
x^+ = \begin{cases} 
  x, & \text{if } x > 0, \\
  0, & \text{otherwise};
\end{cases} \quad (12)
\]

\[
x^- = \begin{cases} 
  -x, & \text{if } x < 0, \\
  0, & \text{otherwise.}
\end{cases} \quad (13)
\]

Now, the problem of updating the bipolar degrees of belief \( \langle c_A^+, c_A^- \rangle \) with the competence dimension of feedback \( F_A(c) \) and \( \langle \sigma_A^+, \sigma_A^- \rangle \) with the sincerity dimension of feedback \( F_A(\sigma) \) may be approached by mapping the bipolar degrees to the single variable \( x = x^+ - x^- \) and compute

\[
y = \begin{cases} 
  x + (1 - x) \cdot f, & \text{if } f \geq 0, \\
  x + (1 + x) \cdot f, & \text{if } f < 0,
\end{cases} \quad (14)
\]

from which \( \langle y^+, y^- \rangle \) may be obtained by means of Equations 12 and 13. Feedback on competence and sincerity of \( s \) about \( d \) will be obtained by

\[
F^d_A(c) = F_A(c) \cdot R(d, \text{con}(A)),
\]

\[
F^d_A(\sigma) = F_A(\sigma) \cdot R(d, \text{con}(A)).
\]

Equations 15 and 16 amount to a projection of either dimension of the feedback along the domains related to the proposed argument, so that only the competence and sincerity about those domains are affected by the feedback.

6 Conclusions

Building on the socio-cognitive model of trust described in [2] and on previous work integrating trust, argumentation and belief revision [3], in this paper we presented a formal framework for modeling how different dimensions of the perceived trustworthiness of the source interact to determine the expected quality of the message, and how deviations from such expectation produce a specific feedback on source trustworthiness. In particular, we characterize competence and sincerity as the key ingredients of trustworthiness, and model both as being domain-dependent. Competence and trustworthiness determine an evaluation of trustworthiness for the source, which in turn produces an expectation on the quality of its arguments. Whenever an argument violates the recipient’s expectation on quality, this produces a feedback on its source—positive if the argument was better than expected, negative if it was worse. Depending on several criteria (most notably, uncertainty reduction and social comparison), this
feedback is distributed among the two key dimensions of trustworthiness, to wit, competence and sincerity. Our model allows to detect the reasons behind the trustworthiness degree assigned to a source in a fine-gained way, and it is useful in such applications where the agents cannot simply avoid the interaction with the untrustworthy sources but they have to reason about trust.

Here, we applied this model to the case of agents exchanging and assessing arguments, but it could easily be extended to the exchange of any kind of factual information. The reason why we focused first on argumentation is because this provides a window on the agent’s reasoning and the resulting process of belief change. This did not play a major role in the present paper, but it would be essential for most extensions of the model: for instance, to study how sources typically exchange information neither randomly nor out of mere kindness, but rather aiming at strategic changes in the recipient’s beliefs and goals, to better serve the source’s own agenda.

However, arguments in this paper were treated basically as black boxes, as it is most often the case in works based on abstract argumentation, in the vein of [6]. This is significant in two respects. First, we did not discuss the two-way relationship between source trustworthiness and trust in the message when what is being communicated is not the argument as a whole, but rather one of its constituents, e.g., a premise, its conclusion, or the inference rule licensing the argument, as in [13]. Finding out that the source is mistaken on the truth of some premise (hence the argument is unsound) rather than on the truth of the inference (hence the argument is invalid) is likely to have very different effects for the feedback on the source, which will have to be investigated in future work. Second, the internal structure of arguments in [3] is limited to deductively valid arguments, again as it is customary in abstract argumentation after Dung [6]. This is a huge idealization with respect to everyday argumentation: as informal logicians and argumentation theorists never tire to repeat [19], we rarely, if ever, exchange deductively valid arguments, while the vast majority of arguments are defeasible, which implies a different sort of consequence relation.

Finally, at present the framework does not capture the cumulative effect of converging sources on argument acceptability, or the effect of converging arguments on belief strength; instead, we use the maximum trustworthiness value among all the sources of an argument, and we do something similar for arguments converging on the same conclusion. This is acceptable in light of all the other issues addressed in this paper, and to comply with length constraints. However, when more than one source offer the same argument or piece of information, its acceptability is positively affected, even if the trustworthiness assigned to the additional sources is not especially high, as discussed in [1]. Conversely, the feedback from the message to the messenger might also depend on how many messengers were making that particular claim, and how much trusted each of them were to start with.

In spite of these limitations, our approach has the important merit of providing a unified framework to represent the effects of source trustworthiness on information assessment, and the converse impact of information quality on
source evaluation. This is the cornerstone of rational trust in communication: we assess neither the messenger nor its message in isolation, but instead capitalize on their mutual interdependence to obtain information on the trustworthiness of both. On this simple foundation, many future studies may (and ought to) build.

References