TRUST IN AGENT SOCIETIES
(TRUST-2011)

14th Edition

Held at:
Autonomous Agents & Multi-Agent Systems Conference
AAMAS 2011
May 2nd, 2011

Taipei, TAIWAN

DESCRIPTION OF THE WORKSHOP

Trust and Trustworthiness (along with related concepts such as privacy, reputation, security, control) have become major research topics in computer science. The multiagent community potentially has a lot to offer, but several conceptual and technical problems must be addressed before it can make practical contributions. Although there is increasing interest in this area within the AAMAS community, this area will need continued support as an affiliated workshop in which are explored new directions and inter-disciplinary interactions so that the AAMAS community maintains a venue for research into trust, reputation, and related topics.

Trust is important in applications such as human-computer interaction to model the relationship between users and their personal assistants. Trust is more than secure communication, e.g., via public key cryptography techniques. For example, the reliability of information about the status of your trade partner has little to do with secure communication. With the growing impact of electronic societies, trust, privacy, and identity become more and more important. Different kinds of trust are needed: 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 many 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 appears to be foundational for 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, coordination, negotiation, with the related trade-off between individual utility and collective interest; or in modeling distributed knowledge and its circulation. In several cases the electronic medium seems to weaken the usual bonds in social control: and the disposition to cheat grows stronger. In experiments of cooperation supported by computers it has been found that people are more leaning to defeat than in face-to-face interaction, and a preliminary direct acquaintance reduces this effect. So, 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.

In the workshop of this edition we will give a special attention about the theme of "TRUST IN SOCIAL COMPUTING". In fact the relationships between social behavior and computational systems are becoming increasingly interwined with interesting bilateral influences. The role of Trust and Reputation has to be deeply analyzed and understood in this new interactional paradigm. We will also call papers coping this theme and we will dedicate a special section of the workshop to this topic.

Just to mention some examples: AI models, BDI models, cognitive models, game theory, and organizational science theories. Suggested topics include, but are not restricted to, the following (here "mechanisms" include considerations of architecture, design, and protocols):

- Models of trust and of its functions
- Models of deception and fraud; approaches for detection and prevention
- Models and mechanisms of reputation
- Role of control and guaranties mechanisms
- Models and mechanisms for privacy and access control
- Models and mechanisms for establishing identities in virtual worlds
- Theoretical aspects, e.g., autonomy, delegation, ownership

- Integration of conventional and agent-based mechanisms
- Policies, interoperability, protocols, ontologies, and standards
- Scalability and distribution across multiple domains or within the global domain
- Test-beds and frameworks for computational trust and reputation models
- Legal aspects
- Trust in Organizations and Institutions
- Application studies (e.g., e-commerce, e-health, e-government)
- **Special Theme:** Trust in social computing

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, TheNetherlands
- 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, Utrecht University,The Netherlands
- Eugen STAAB - Computer Science, Imc AG, Germany

CONTENTS

A Unified Framework for Trust in Composite Networks

Sibel Adalı[1], William A. Wallace[1], Yi Qian[1],
Priyankaa Vijayakumar[1], and Munindar P. Singh[2]

[1] Rensselaer Polytechnic Institute
{adalis, wallaw, vijayp, qiany3}@rpi.edu
[2] North Carolina State University
singh@ncsu.edu

Abstract. A composite network is one that captures participants (represented by vertices) and relationships (captured by edges) at multiple levels of abstraction in a cohesive manner. Of special interest are composite networks that include (1) social networks whose participants are people and relationships are human relationships; (2) information networks whose participants are information resources and relationships are those of flow and reference; and (3) communication networks whose participants are network elements such as routers and relationships are those of connectivity. It is well-recognized that the concept of trust potentially applies in all kinds of networks where the participants carry some level of autonomy or decision making and the relationships include those of dependence and risk. We seek in this paper to initiate a systematic treatment of trust in composite networks. We provide a general architecture and show how it may be instantiated computationally.

1 Introduction

The purpose of this paper is to broaden our understanding of trust by considering its function in decision-making by agents in networks. A review of the literature shows that trust arises in a variety of network settings, ranging from social relationships to computer protocols. An agent may have to interact *within* and *with* many of these networks in any given decision situation. Therefore, any model of trust for decision making in networks must consider the unique characteristics of each of the networks involved in a decision situation.

In this paper, we address the problem of modeling trust in the context of decision-making in composite networks. To illustrate such a network, let us consider a person, the *trustor*, who needs to rely upon another person, the *trustee*, in order to make a decision. In addition, the trustor queries electronic sources to obtain information to support the given decision-making activity. The sources of information and their relationships to each other form the information network. Finally, all of these interactions are mediated by computer systems that store and transmit information between these sources and agents in the network. The communication channels as well as the entry points to the information network constitute the communication network that must be traversed to accomplish a task, such as sending information. In this manner, the above simple scenario involves the *social*, *information*, and *communication* networks. Pulling these three networks together into a *composite network*, we consider the problem of how trust may be modeled and computed in a composite network.

A decision concerning trust might require one to traverse paths in this composite network that span nodes and edges of multiple types. For example, person A issues an order to person Z based on information that person A obtained from source B over a communication channel C. Should person Z follow this order? How much should person Z trust person A's ability to process the information he or she receives from source B? How trustworthy are the information source B and the communication channel C? Ultimately, the evaluation of trust by person Z must take into account all of these factors before person Z can rationally decide to follow the order.

We first describe the trust problem in general and then introduce a model for trust evaluation in each individual network type. We then introduce a framework for implementing a unified model over all of these network types and give examples of each component in our framework. Our aim is to show how a system that computes trust in the composite network can be implemented and eventually used as the basis for a unifying set of measures and metrics of trust. Whereas a great deal of research in trust has gone into modeling of trust, there is no research that explicitly addresses trust in composite networks.

2 Modeling Trust

Our aim in this paper is to model how much a person or computational agent trusts another in a special decision context. The actor is faced with a situation where she has to make a decision and this decision requires trust. This could be that trusting another person is needed to accomplish a task. It could also be that it is necessary to decide whether to trust information from a specific source or not. In some instances, the actor has to decide whether a system can be trusted to transfer information from one location to another. These decisions may not be independent: social interactions might take place over a network, rely on information exchanges with third parties, and so on. In fact, in most cases the decision are interdependent. Hence, the trust decision is a composite of these individual components. The associated trust model must be capable of incorporating these interactions. In our model, we consider trust as a directed edge between two entities, a trustor and a trustee. The trustor is a cognitive agent. Trustor can trust (as a trustee) a person, an information source or a system. In the following, we seek to define the cognitive model for each type of trustor and show how we can combine all of them in a single framework.

In a decision context, the trust relation is defined with respect to the expectations of the trustor (A): whether the trustee (B) will act a certain way, whether the information provided by B is correct, or whether B will transmit information correctly. In order for trust to be relevant in a particular decision situation, uncertainty and vulnerability must be present in the given context—that is, the trustor must deal with risk. Further, the trustor has a need that has to be satisfied and the trustee has the potential to satisfy this need. We refer to this as the dependence of the trustor on the trustee. In essence, dependence and vulnerability define what the trustor potentially has to gain if the trustee provides help and to lose if no help is provided or trust is misplaced. There could be uncertainty in many aspects of the context. These three components of context together are treated as a precondition of trust.

The decision context not only specifies these preconditions, but also incorporates many other pieces of information. For example, the goals of the agent and his or her mission and motivations (broadly understood) are part of context. Resources available to the agent such as availability of resources and skills of team members can also be considered a component of the context. Finally, context also incorporates hard and soft constraints that limit the scope of actions possible: physical limits, time limits, battery power, norms introduced by an organizational hierarchy are all limits on what actions are possible and should be considered. Cognitive resources (for example, the abilities to remember and reason) are resources but also introduce limitations. These different components of context together determine to a large degree how trust is computed.

We represent trust as a cognitive model operating within the mind of the trustor. We note that the trustor is not necessarily a rational decision maker. The model allows for more automatic, emotional, stress-influenced decision making as we discuss in the next section.

2.1 Social Trust

We first describe the notion of social trust. The social trust in our context corresponds to trusting another person to accomplish a task. Hence, trust in this context refers to the degree which the trustor considers the trustee to be capable and willing to accomplish this task. The model is shown in Figure 1 (figure taken from [1]). First, the trustor identifies with the trustee based on perceptual signals such as the look and feel of the trustee. Research in cognitive science [2] shows that facial features are processed much faster than specific information about the person, and may have a significant influence in the final trust evaluation when quick decisions are needed. The second component of trust is the trustworthiness of the trustee. Note that the trustworthiness of a trustee is estimated by the trustor based on what the trustor knows about the trustee. The trustworthiness includes evaluations on the competence of trustee (that is, the ability to accomplish the necessary task), positive intentions, ethics, and predictability. In essence, predictability influences the uncertainty involved in the trust evaluation. The information used to infer trustworthiness can be based on the direct experiences the trustor has had with this specific trustee in this context as well as the indirect information obtained from the social network through recommendations and queries. Other components of trust are based on social aspects of trust and the propensity to trust. We consider propensity an attribute of the trustor, whereas the social trust defines attributes of the trustee. Both of these attributes are used by the trustor to evaluate trust.

Social trust can be considered attributes of the trustee that drive their meaning from the social network, through norms and culture. For example, the trustor might trust another for a health question because he is a doctor. This is because the role "doctor" has an associated certification which carries a special meaning through the social norms. Similarly, culturally defined roles such as parent also fall in this category. Further, social science research considers social ties based on the similarity or complementarity of actor's attributes (assortativity), relational mechanisms such as reciprocity, repetition of interactions, degree, network location and proximity-based mechanisms that have to do with foci that bring actors together [3].

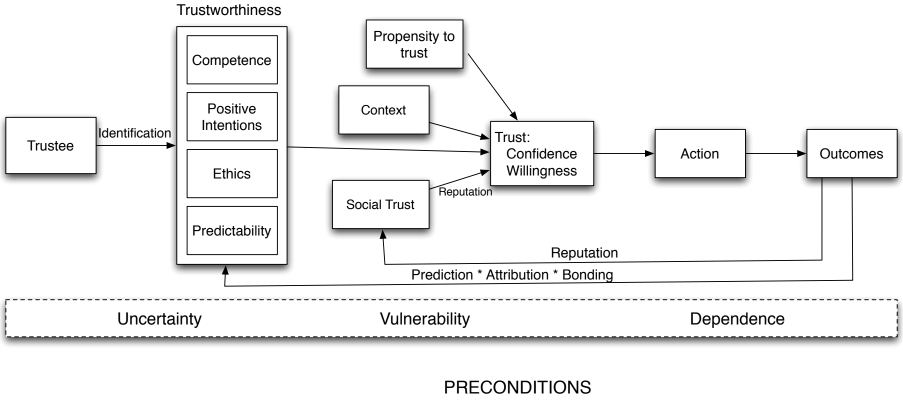


Fig. 1. Social trust.

In our model, the trustor is influenced by all these factors in evaluating trust: knowledge of the trustee is evaluated through perceptual cues, social cues, and past experiences obtained through direct or indirect experiences. All these factors are evaluated within the given context and the trustor's propensity to trust to construct an evaluation of trust. For example, if a probabilistic interpretation of trust is used, then the trust computation may return both a value (an expectation) and uncertainty of this value. This could be modeled, for example, by a probability distribution.

At this point in the process the trustor has made his/her trust assessment and now has to use it in taking an action (making a decision). This requires the trustor to have confidence that the trust assessment will lead to "good" consequences and be willing to take the risk by making a decision. This means that the individual differences determine the appropriate threshold of trust the trustor needs to take an action. For example, if trust is high but trustor's self confidence needed to take an action is low, the trustor may not be willing to take an action.

Whenever an action is taken and the trustor has the ability to observe the results of this action, then the trustor can incorporate new information about the trustee into his or her knowledge base through processes of prediction, attribution and bonding. The social network is also impacted by the actions which in turn influence the computation of the social factors that relate to trust.

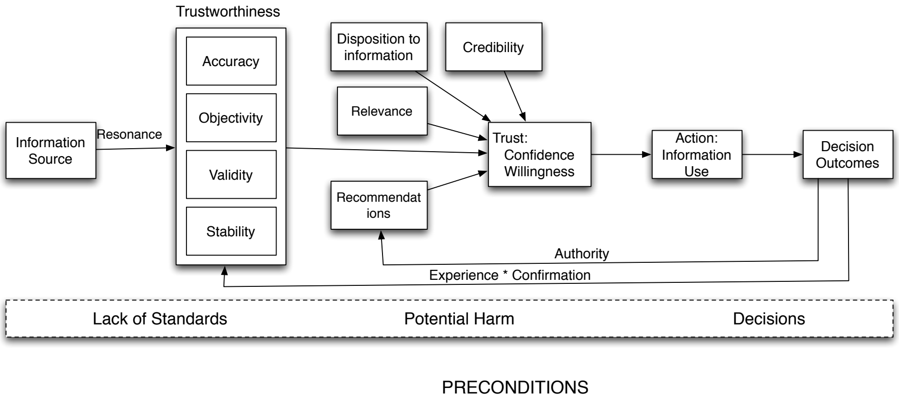


Fig. 2. Trusting information.

2.2 Trusting Information

We now define the model for evaluating how to trust information from a specific source. If the source is not known, then trust for information is not meaningful. Only, the credibility of the information can be considered. The information source in our model could be a person or any entity that provides a way to store and display information. Clearly, these entities provide certain policies that determine what content could be posted and when, which in turn impact their trustworthiness. When determining how trustworthy a source is (that is, for providing trustworthy information), different considerations than the social trust model are used. Our model is shown in Figure 2. Note that resonance is the mechanisms used to process perceptual information such as user interfaces [4], whether the presentation of the information resonates with the trustor. The trustworthiness of a source is based on its accuracy, objectivity, validity, and stability. In this case, stability is analogous to predictability, determining the confidence the trustor places in the trust evaluation. Social trust mechanisms relate to the degree that this source is considered authoritative in this context, defined solely by the social norms that assign this source an authority on top of measure of trustworthiness. For example, a commander may be considered an authoritative source of information even though his/her trustworthiness is in question.

Unlike social trust, a secondary evaluation of the information credibility is used to determine to which degree the information can be trusted. Hence, the trust has a component based on its source (that is, the trustee) and another component based on its credibility evaluated by the trustor. The actions in this instance involve the use of the information to make a decision. Experience and confirmation are mechanisms used to update the information about the trustee obtained through actions and other means from the outside world. Note that the vulnerability in the information context refers to the potential harm when the information is not correct. The lack of standards create both uncertainty and dependence on the trustee for the required information.

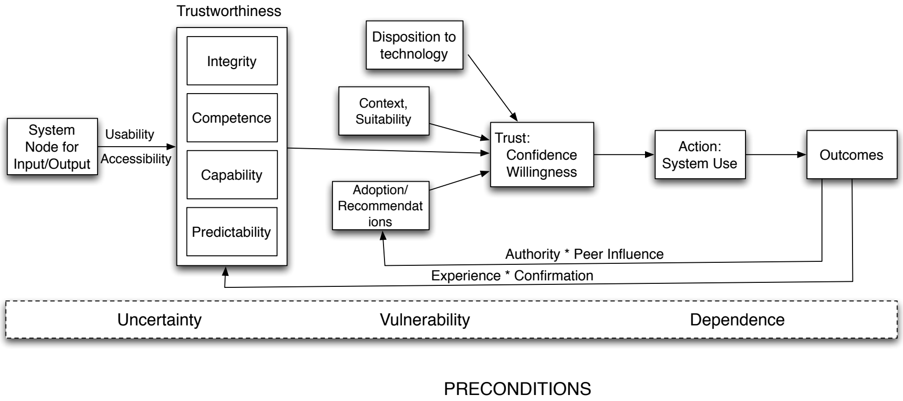


Fig. 3. Trusting a system.

2.3 Trusting Systems

The third type of trust we consider in a decision making context is the decision to use a system to either transmit or store information to another party. There is a great deal of work in this area [8–10]. Similar to information, the user interface to the systems play a role in the trust to use them. Some of the components of trustworthiness have to do with the capability of the system to store and transmit data without releasing it to parties (not intended by the trustor to receive it), capability with respect to nonrepudiation, authentication, and confidentiality. We summarize these aspects as integrity. Note that in our model, we are interested in the evaluation of the trustee on this aspect by the trustor based on the trustor's available knowledge. Similarly, competence is used to refer to the ability of the system to accomplish a task. It is especially meaningful when the trustee incorporates a cognitive agent. For example, a system that is capable only of transmitting noisy data is not highly competent. Additionally, the precision and recall of the agent for retrieval tasks, skill in automation tasks [5, 7] are all relevant factors for competence. Predictability refers to how available the system is and how much the quality of service varies over time, in essence how stable the system is. Finally, capability refers to a combination of factors like bandwidth, capacity, processing power, and so on. The capability determines how fast and how much of trustor's data can be transmitted through the system. As with all components of trustworthiness, these factors can be evaluated based on direct experiences with the system or through indirect information obtained through recommendations from one's social network. At the social level, we can consider the adoption of the system as a factor, as the trustor is likely to use a system that is likely to reach the needed individuals.

3 Toward a Unified Model

Based on the foregoing motivations, we now outline what it would take to develop a unified model of trust for composite networks. Our proposal has two main components: a software architecture and a computational approach.

3.1 Architecture of a Trusting Agent

Figure 4 outlines our proposed architecture in conceptual terms. We imagine that an agent deals with an outside world or environment and finds itself in a particular decision context.

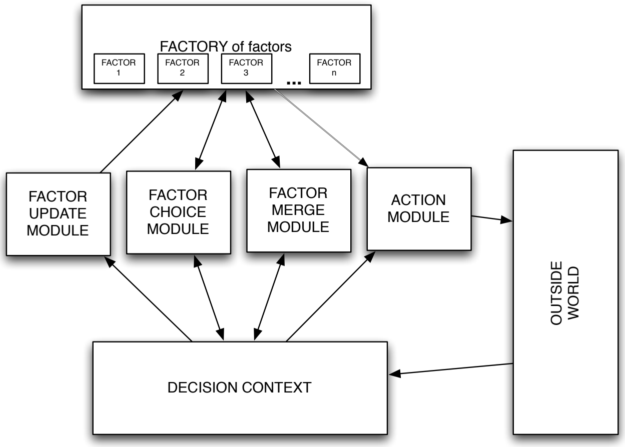


Fig. 4. The architecture of the unified system in conceptual terms.

Available to the agent are a range of possible criteria or *factors* of trust. One can think of these as being treated in a software engineering approach via a factory. That is, each factor can be thought of as instantiating a particular interface, which supports prespecified generic methods such as to initialize, read, and update. Specifying such an interface modularizes and streamlines the creation of a reasoning system for trust. Figure 5 illustrates how an agent may maintain and exploit its knowledge regarding each factor of interest.

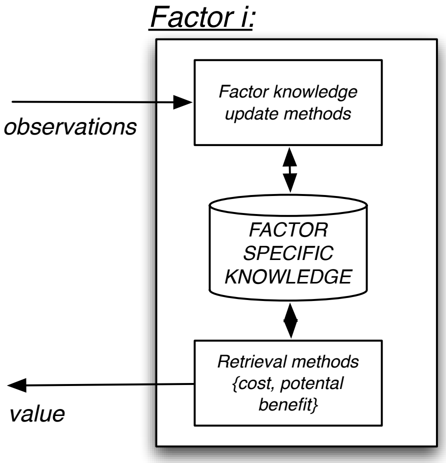

Factor i:
observations
Factor knowledge update methods
FACTOR SPECIFIC KNOWLEDGE
Retrieval methods {cost, potental benefit}
value

Fig. 5. Schematic of factors involved in placing trust.

Back in Figure 4, the principal decision-making of an agent is driven by what factors it selects as relevant and how it combines the trust assessments generated by the selected factors. These feed into the action module, which determines whether and what action to take. The decision of the agent is a true decision in that there is always at least a pair of alternatives from among which the agent has to choose. If there were no choice to be made, the whole exercise of determining trust would be moot.

The decision and action by the agent have consequences in the world, affecting the world in some manner, and through the world indirectly affecting the agent's local outcomes as well. For example, if we adopt a simple reinforcement learning style model, the agent may be rewarded or penalized by the world based on the action it takes. If we adopt a richer cognitive-emotional model, the action of the agent could have consequences in terms of the happiness, disappointment, betrayal, or other such cognitive-emotional attitude that the agent may feel [11].

The upshot of closing the loop in the above manner is that it gives the agent an opportunity to learn from its experience. The learning is placed in the factor update module, whereby the agent adjusts its estimates of the effectiveness and value of the factors it had selected in the previous episode.

3.2 Representing Trust

We imagine that many representations are possible for the wide ranging criteria that we have motivated for a unified treatment of trust. However, for concreteness, we consider one possible representation—as a way to suggest how the above-mentioned architecture could potentially be realized computationally.

So as to be able to treat a varying set of factors in a modular manner, a natural approach is one based on probability theory. That is, given a specific factor, a trustor can express a probability distribution for the trustee being trustworthy in the present decision context given a specific value for that factor. We can then develop a suitable calculus for merging factors and for updating the distributions for various factors. Reasoning with distributions directly can be complex.

Instead, we adopt an approach originated by Jøsang [12] and enhanced by Wang and Singh [13] (the main enhancement by Wang and Singh is that their approach can handle conflict in evidence correctly as reducing certainty whereas Jøsang's approach disregards conflict). In this approach, we can represent each factor's impact on the trustworthiness of a trustee in terms of the mean probability and a measure of certainty or confidence in that probability. The approach supports a natural calculus by which to combine the impacts of two or more factors, each potentially weighted differently.

Wang et al. [14] recently have shown how to update the predictions made by the above approach in light of evidence, in a manner that takes into account the relative polarity and strength of the prediction and the new evidence.

To summarize, we can see that our general architecture for trust can be realized, if in a simple manner, by a probabilistic approach. We expect that other realizations would be needed that would, on the one hand, be more sophisticated in their treatment of the cognitive and social concepts involved and, on the other hand, be more sophisticated in their treatment of utilities and economic preferences.

4 Discussion

We now discuss our approach in a broader setting.

4.1 Literature

One of the best known cognitivist approaches to trust is that of Castelfranchi and Falcone [17], who understand the *trustor* as trusting the *trustee* based on their respective beliefs and intentions regarding the plans of the trustor, the (apparent) willingness and ability of the trustee to support such plans, and the explicit reliance of the trustor on the trustee for accomplishing said plans. The present approach has not pursued plans to any detail but we imagine is broadly compatible with a plan-based account.

Further, the notion of relational capital as articulated by Falcone and Castelfranchi [18] is also relevant to our approach. It particularly applies to the setting of social networks wherein the reputation gained by a person in being deemed trustworthy by others can potentially be parlayed into its obtaining trustworthy behavior from others. Interestingly, the notion extends naturally to information and communication networks as well. For example, an information node that had provided high-quality information to others might expect to be rewarded by high-quality information from others. And, a communications node that had diligently forwarded packets on behalf of other nodes might expect that others would forward packets on its behalf. More generally, the foregoing ideas relate to the setting where the parties involved (whether they be humans or information or communication resource nodes acting as surrogates for humans) are considered rational and strategic. In such a setting, agents must gain from being trustworthy or they would have every motivation to defect against the others. Hazard and Singh [19] have studied some of the technical aspects of such a model, especially in terms of potential axioms for trust and a result mapping a rational agent's trustworthiness to its patience for long-term gain.

Works by Barber and Kim [15] and Fullam and Barber [16] have studied rich models by which an agent may update the trust it places in another. This work has mostly been focused on the information network (in our terminology), because it is concerned with judging the trustworthiness on an information source and updating estimates of such trustworthiness. However, the richness of the model suggests a potential for application in our factor-based architecture.

Recently, Johnson et al. [20] have examined the idea of social interdependence as underlying teamwork. The emphasis on interdependence is crucial as a basis for trust in our purposes. Johnson et al.'s applies primarily at the level of social networks. It would be interesting to elaborate it in connection with composite networks.

4.2 Conclusions and Directions

Although trust has long been studied in connection with networks, the study of trust with reference to composite networks offers new opportunities and challenges for research. We have only recently initiated this effort.

Some themes of particular interest are the following. It would be instructive to revisit the idea of the propagation of trust. There is some natural intuition that trust can propagate in that if A trusts B and B trusts C, A can be expected under certain reasonable assumptions to trust C [22]. Indeed, the value of referrals in human networks for ages as well as in modern business networks relies upon an inherent ability to propagate trust. However, such propagation is far from trivial or obvious, as Falcone and Castelfranchi [21] have recently argued. An outstanding challenge is to identify the assumptions under which trust can propagate transitively across the three main networks that we consider, and to find ways to compute it effectively when it can propagate.

Another important challenge is to incorporate the notions of rationality and utility into the fold. This work would build on studies such as that of Hazard and Singh [19] but expanded to account for the additional structural properties of the three networks of interest, and of the resulting composite network.

A deep study of how social norms contribute to trust and how this notion of social trust can be inferred from network structure, network flows and the evolution of the network structure is a topic of future study.

Acknowledgments

This research is supported through the authors' participation in the Network Science Collaborative Technology Alliance (NS-CTA) sponsored by the U.S. Army Research Laboratory under Agreement Number W911NF-09-2-0053.

The content of this paper does not necessarily reflect the position or policy of the U.S. Government, no official endorsement should be inferred or implied.

We would like to thank Jin-Hee Cho and Gerald Powell, ARL, Jennifer Mangels, CUNY and anonymous reviewers for their insightful comments.

References

1. K. Kelton, K. Fleischmann and W.A. Wallace (2000) "Trust in digital information". Journal of the American Society for Information Science vol. 59 (3) pp. 363-374
2. J. D. Rudoy and K. A. Paller. "Who can you trust? Behavioral and neural differences between perceptual and memory-based influences". Frontiers in Human Neuroscience (2009) vol. 3
3. M. T. Rivera, S. B. Soderstrom and B. Uzzi "Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms". Annual Review of Sociology (2010) vol. 36 pp. 91-115
4. B.J. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon and P. Swani "What makes Web sites credible?: a report on a large quantitative study". Proceedings of the SIGCHI conference on Human factors in computing systems (2001) pp. 61-68
5. J. Lee and N. Moray. "Trust, control strategies and allocation of function in human-machine systems". Ergonomics (1992) vol. 35 (10) pp. 1243-1270
6. L. J. Chang, B. B. Doll, M. v. Wout, M. J. Frank, A. G. Sanfey "Seeing is believing: Trustworthiness as a dynamic belief", Cognitive Psychology 61 (2010) 87-105
7. S. M. Merritt "Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions", Human Factors, Vol. 50, No. 2, April 2008
8. C. L. Corritorea, B. Krachera and S. Wiedenbeck "On-line trust: concepts, evolving themes, a model" Int. J. Human-Computer Studies 58 (2003) 737-758.
9. D. Artz and Y. Gil. "A survey of trust in computer science and the semantic web". Web Semantics: Science, Services and Agents on the World Wide Web (2007) vol. 5 (2) pp. 58-71
10. T. Grandison and M. Sloman. "A survey of trust in internet applications". IEEE Communications Surveys and Tutorials (2000) vol. 3 (4) pp. 2-16
11. Clark Elliott "The Affective Reasoner: A Process Model of Emotions in a Multi-agent System". Ph.D. Dissertation, Northwestern University, 1992
12. A. Jøsang "A subjective metric of authentication". *Proceedings of ESORICS*, LNCS 1485, Springer, 1998, 329–344
13. Yonghong Wang and Munindar P. Singh "Evidence-Based Trust: A Mathematical Model Geared for Multiagent Systems". ACM Transactions on Autonomous and Adaptive Systems (TAAS), volume 5, number 4, November 2010, 14:1–14:28
14. Yonghong Wang, Chung-Wei Hang, and Munindar P. Singh "A Probabilistic Approach for Maintaining Trust Based on Evidence". Journal of Artificial Intelligence Research, volume 40, January 2011,221–267

15. K. S. Barber and J. Kim. Belief revision process based on trust: Agents evaluating reputation of information sources. In R. Falcone, M. P. Singh, and Y.-H. Tan, editors, *Trust in Cyber-Societies*, *LNAI* 2246, pp. 73–82, 2001. Springer.
16. K. Fullam and K. S. Barber. Dynamically learning sources of trust information. *AAMAS*, pp. 1062–1069, May 2007.
17. Cristiano Castelfranchi and Rino Falcone "Principles of trust for MAS: cognitive anatomy, social importance, and quantification". Proceedings of the 3rd International Conference on Multiagent Systems (ICMAS), IEEE Computer Society, 1998, 72–79
18. Cristiano Castelfranchi, Rino Falcone, and Francesca Marzo "Being Trusted in a Social Network: Trust as Relational Capital". Trust Management: Proceedings of the iTrust Workshop, LNCS 3986, Springer, Berlin, 2006, 19–32
19. Christopher J. Hazard and Munindar P. Singh "Intertemporal Discount Factors as a Measure of Trustworthiness in Electronic Commerce". IEEE Transactions on Knowledge and Data Engineering, 2010, In press
20. M. Johnson, J. M. Bradshaw, P. Feltovich, C. Jonker, M. B. van Riemsdijk, M. Sierhuis. Coactive design. *AAMAS COIN Workshop*, pp. 49–56, 2010.
21. Rino Falcone and Cristiano Castelfranchi "Trust and Transitivity: A Complex Deceptive Relationship". Proceedings of the 12th AAMAS Workshop on Trust in Agent Societies (Trust), 2010, 43–54
22. Chung-Wei Hang, Yonghong Wang, and Munindar P. Singh "Operators for Propagating Trust and their Evaluation in Social Networks". Proceedings of the 8th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), Budapest, May 2009, 1025–1032

Unbiased Trust Estimation in Content-Oriented Social Networks

Suratna Budalakoti, David DeAngelis, and K. Suzanne Barber

The University of Texas at Austin
The Laboratory for Intelligent Processes and Systems
University Station C5000, ACE 5.124
Austin, Texas, 78712-0321 USA
{sbudalakoti,dave,barber}@lips.utexas.edu

Abstract. Online communities based on user generated content (UGC) often rely on a small set of highly loyal and productive users to identify or create content that would interest their broader audience. Through continual contact, this user base often develops informal reputational and social ties among themselves. Websites often encourage the formalization of these trust networks by providing various social networking features, to increase engagement and loyalty. A user may come to trust a subset of the others as a consistent source of good content, while avoiding the remainder, leading to influence-based fragmentation. We discuss the impact that these emergent social behaviors can have on the quality of reputation scores, and develop algorithms that are able to take them into account while calculating user reputation.

Keywords: trust, reputation, social network, user-generated content

1 Introduction

Websites built on user generated content (UGC) are prevalent, and the perceived value of this content is growing rapidly [18]. Sites like Twitter, Yelp, Digg, Reddit, eBay, Yahoo! Answers, Amazon, and many others rely on content which is created by their users, whether it is product reviews and descriptions, restaurant suggestions, movie recommendations, or any other kind of information. Often these websites allow each user to create an online identity. Through contributions to the site, users build a reputation through the collective whole of other users. This reputation and its associated measure of trust form a social network.

Social networks can be explicitly defined using friendship or linking mechanisms, or they can be implicitly created, by users simply tracking the identity of a content creator. UGC websites often encourage users to formalize these networks of trust by providing various social networking features (contacts, follows, and friendships). These features are designed to increase engagement and loyalty in this active user base and to encourage the growth of the base in the long term. Due to these reasons, UGC forums are also under online social networks (OSNs) as *knowledge-sharing oriented online social networks*[5, 13] or content oriented social networks (COSNs). In the three primary activities users perform on OSNs, authoring content, viewing content, and networking, COSNs are OSNs where the emphasis is on authoring and viewing content. This is in contrast to *networking oriented social networks* (NOSNs) such as Facebook, which are driven by the users' social relationships and networking activity. Thus, on a networking oriented OSN, users will be most interested in information about their close friends, while in a knowledge oriented OSN, a piece of information may have intrinsic value (depending on its quality, relevance, etc.), independent of which member introduces it to the group.

Often such UGC sites rely on a small set of highly loyal and productive users whose actions interest the broader audience. Such users are the most trustworthy users. Historically trust is defined as a measure of the truthfulness or reliability of an agent [7]. In this research the most trustworthy agents are the users whose contributions to the community add the most value. The social networks on these sites, while helpful in increasing user engagement and allowing core users to quickly find information from sources they trust, can be problematic. The formation of social networks can give rise to various social phenomena such as nepotism, reciprocity, and cyber-balkanization [19], which can distort the rating processes of the core set of users.

These emergent social network phenomena impact the accuracy of users' reputations. The contributions of this research are a characterization of the impact of social network phenomena on user reputation and an algorithm for identifying the most trustworthy users in a UGC-based community.

2 Related Work

One source of trust, apart from personal experience, is reputation. Reputation is an aggregate indicator of the trustworthiness of an agent, as observed by other agent. A good reputation score implies that an agent, or user, is generally believed to be trustworthy. Barber and Kim explore this process of belief revision based on this type of reputation in in [1]. Online communities, particularly general UGC websites, often have a large, sparsely connected user base. The likelihood of one user A interacting with another particular user B in a large system is very small, and often multiple interactions are necessary to develop an accurate model of direct trust. Therefore it is impractical to rely on direct interaction for a large part of their user base. Instead, such website rely on an aggregated reputation model from the community as a whole, or "neighborhood reputation" [17], to identify valuable contributors.

Lerman *et al.* have investigated the spread of content in Digg and Twitter and discovered that the most prolific users find and consume content through their social networks [11]. These users also provide the most trust rating for reputation aggregation, hence their behavior patterns can be extremely influential. Lerman's work highlights the importance of social influence in governing which content is promoted in UGC websites. Similar findings have caused Digg to implement a policy where they discount endorsements, “Diggs”, by users who are in the same social network as the original poster [16]. We argue that while social influence causes users to vote for their connections, the opposite effect of homophily-based selection [3] needs to be taken into account as well. It is possible that users add others to their social networks because they like the content that they produce. In that case, discounting all votes from a user’s social network connections could be misleading. Instead, we propose an approach based on estimating the user’s intent behind his/her votes, instead of simply discounting all social contact based votes. In [4], Ghosh and Lerman show that it is possible to predict which content will flourish by examining the flow of that content through the social network in its early stages. Alternatively, this research focuses on identifying valuable contributors.

Building and trusting in others on an anonymous Internet is difficult. Often there is little consequence for antisocial behavior and users behave in a greedy manner. According to Resnick *et al.* effective trust and reputation models require that entities are long-lived, feedback about current interactions is captured and distributed, and past feedback guides buyer decisions [15]. UGC websites have the necessary infrastructure to address these points and build meaningful reputation models. Users on UGC websites have a persistent identity (their user names), and trust is established over time by observing their actions. Another fortunate benefit of building trust models of users in UGC communities is the centralized nature of UGC websites. In contrast to decentralized trust models as proposed by Yolum and Singh [21], the website infrastructure of a UGC community monitors and aggregates every interaction. Constructing reputation in this centralized fashion allows all users to access the same reputation information for guiding their decisions.

According to Pavlov *et al.* a potential pitfall of reputation information is that it may be provided in a strategic manner for numerous reasons including reciprocation and retaliation [14]. Social networks suffer from a similar problem. Phenomena between users such as nepotism, reciprocity, and retaliation can distort common measures of trustworthiness. To address this problem, we propose a mixture-model based approach which explicitly models the behavioral aspects of interactions on a COSN as a component. The other component of the mixture is expected to model the process by which users assign unbiased ratings to high quality content. By estimating, for each user, the likelihood of their behavior belonging to either component, the algorithm attempts to identify users who indulge least in behavioral patterns that may mislead a reputation system.

3 Background

The two key roles on a UGC forum (or COSN) are that of content creators, users who create content with the expectation that it may interest others, and users who consume the content. The two roles are not mutually exclusive, the same user may be a content creator or consumer at different times, depending on the circumstances. There are many ways in which consumers may express their opinion of a piece of content: the very fact that the consumer accessed the content (by say, clicking on a link to it) can be seen as a positive affirmation. Also, many website provide ways by which users can express their approval, for example, by ‘upvote’ links for users to click. We call any such action by which a content consumer may express their approval of content as a *selection*.

A selection can be seen as a vote of confidence in the quality of content produced by the content creator, and the total count of selections can provide a good initial estimate of the quality of a user’s content, or a user’s reputation. One problem with this approach is that all users are not equally good judges of quality: some users may be more qualified, or they may be more involved in the forum, and thus have a better understanding of the goals and ‘personality’ of the forum. A more advanced approach would be to weigh selections by some measure of the selecting users’ reputation in the forum, an approach that might yield an algorithm similar to eigentrust [8], proposed for reputation estimation in peer-to-peer networks.

Another important problem is that, even in the case of users who may be highly reputed on a forum, the motivations behind a selection they made is not always clear. The reason for this is the social aspect of UGC forums: over time users develop social relationships with other users, and these relationships impact choices about the content they consume or favor. As UGC forums rely on these users to select content of general interest, incorporating these biases while identifying content of general interest can adversely affect the selection quality. Some documented examples of such biases are:

1. *Reciprocity*: A common social norm on many forums is for users to provide a reciprocal link in response to a link. This norm can be seen as a form of courtesy, but is also exploited by some users to increase their link count. Reciprocity of links is a well-documented phenomenon on the websites Flickr [10] and Twitter [20].
2. *Social Voting*: Many content-sharing sites such as Digg and Yahoo! Answers allow users to add other users as contacts or friends. The aim is to increase engagement: the site is designed so that users find it easy to get updates on the activities of their contacts. A side-effect is that since users find interesting stories via their contacts, users with many contacts find it much easier to promote their content. Social voting has been documented on the website Digg [4] as well as Flickr [12].

In other words, the reputation that users aggregate over time does not depend only on their quality, but also on many behavioral side-effects of their social network interactions. In this paper, we propose a mixture model that assumes that a user’s rating behavior could be driven by one of two intents/motivations:

1. *Content Quality*: The responder’s expertise in a topic determines the quality of the content produced by him/her. A selection based on content quality recognizes the content creator’s authority, and should considered when estimating his/her reputation.

2. *Social Affinity*: The social affinity between a content creator and producer is independent of the content quality, and depends on their relationship with each other, which may be observed or modeled, given information about their online social network links.

4 User Reputation

A natural way to define a user's reputation in a UGC forum is the number of times content created by him/her has been selected, or rated positively by another user. A more sophisticated approach would be to weigh each selection by the reputation of the user making that selection.

Then, let the reputation (or authority) of user i in a topic be written as r_i, and the number of times user j selected content by user i, r_{ji}. Let N_S be the total number of selections made. Then r_i can be written as follows:

$$r_i = \sum_{j=1}^{N} r_j \frac{r_{ji}}{N_S} \tag{1}$$

where N is the number of users. Now, let the number of rating by user j be written as q_j. Then, after normalizing with the total reputation of all users in the system, we can rewrite r_i as follows:

$$r_i = \frac{\sum_{j=1}^{N} q_j \cdot r_j \cdot p_{ji}}{\sum_{j=1}^{N} q_j \cdot r_j} \tag{2}$$

where p_{ji} is the fraction of questions by j answered by i. Dividing both numerator and denominator by N_S, we get:

$$r_i = \frac{\sum_{j=1}^{N} \frac{r_j}{N_S} \cdot r_j \cdot p_{ji}}{\sum_{j=1}^{N} \frac{q_j}{N_S} \cdot r_j} \tag{3}$$

Interpreting $\frac{q_j}{N_S}$ as the probability that user j will provide a rating, written as P_j^q, we get:

$$r_i = \frac{\sum_{j=1}^{N} P_j^q \cdot r_j \cdot p_{ji}}{\sum_{j=1}^{N} P_j^q \cdot r_j} \tag{4}$$

4.1 Absorbing Random Walk Interpretation

This can be written in matrix form: let Q be a diagonal matrix, where $Q(i,i) = P_i^q$, let P be a matrix such that $P(i,j) = p_{ij}$, and let r be a vector corresponding to $r_{i \ldots N}$ above. Then the above equation can be written as:

$$(QP)^T r = r \tag{5}$$

We can add a small uniform prior probability matrix ez^T to P, where $e_i = 1$ for all i, and z sums to 1. This signifies a small probability that any user can select any other user, even with no current evidence in the data. Adding 1 to the denominators of $Q(i,i)$ preserves a probabilistic interpretation. A restriction that r sum to 1 can be added. Then we can rewrite the above equation as:

$$(QP + ez^T)^T r = r \tag{6}$$

Solving this gives

$$r = (I - QP)^{-T} z \tag{7}$$

Let $T = QP$. Then $r = (I - T)^{-T} z$[1]. As all rows of matrix T sum to less than 1, we can interpret T as the transition matrix for a reducible Markov chain with $N + 1$ states by adding an extra recurrent absorbing state, which is the exit state. At any timestep, if the system is currently in state i, it transitions to the exit state with a probability $1 - Q(i,i)$, and to another state j with probability $Q(i,i) \times P_{ij}$. Then $R = (I - T)^{-1}$ is the definition of fundamental matrix of an absorbing Markov chain, that is $R = I + \sum_{i=1}^{\infty} T^i$ [9]. So, if a random walk is executed across the absorbing chain, R_{ij} is the expected number of visits to state j before exit, if the walk started in state i. As z is a probability vector, $r = R^T z$ gives the expected time spent in each state, if the initialization probability of the walk at vertex i is given by z_i.

4.2 Relationship to Eigentrust

Pagerank[2] is a popular algorithm for link analysis over a collection of hyperlinked documents. A variation of pagerank, called eigentrust [8], was proposed by Kamvar *et al.* to estimate user reputation in P2P networks. Applying the eigentrust formulation to our problem would define user reputation as:

$$((1-c)P + cez^T)^T r = r \tag{8}$$

where c is a parameter, called the teleportation probability, and usually set to 0.85. The common approach to solving this equation is via an iterative method. However, solving algebraically, as described in [6] gives:

$$\Rightarrow r' = (1-c)(I - cP)^{-T} z \tag{9}$$

Comparing equations (7) and (9), we see that r and r' differ only by a constant, $(1 - c)$. So (7) provides a generalization of the pagerank vector: $r = r'$ when $Q(i,i) = c$ for all i, that is, when all users are weighed equally, irrespective of the number of ratings provided.

One intuitive interpretation of pagerank in the context of webpages is the random surfer model: intuitively, a webpage's authority is estimated as the probability that a random web surfer would visit the page given that he/she starts at

[1] In practice, the inverse need not be calculated, but r can be calculated by solving the set of equations using Gaussian elimination, based on T and z.

a random page, and selects a random outlink at each timestep. The vector r can be understood in terms of the random surfer model as follows: in pagerank, there is a constant probability c with which a surfer gets bored at any timestep and teleports to another random page. This seems reasonable for webpages, where the number of links provided may have little relationship to the quality of the page, but for UGC forums, more active raters are more likely to be seriously interested in the forum, and likely to be better judges of content quality. In our formulation, the probability of random teleportation varies inversely with the number of ratings provided by the user. It would be useful to have this effect level off at some point, so that users cannot increase their influence as questioners simply by asking a lot of questions. For this reason we use a sigmoid function to set Q. We set $Q(i,i) = \frac{1}{1+e^{-0.05q_j}}$. This means that for questioners who have provided 100 or more ratings, $Q(i,i)$ is effectively equal to 1.

5 Mixture Model Based Reputation Estimation

5.1 Algorithm Outline

Raters' fairness or objectivity is supposed to estimate the degree to which their ratings are motivated by the quality of the content rated, as opposed to the influence their social network has on them. We define a hidden variable vector *objectivity* o, where o_i as a measure of the degree to which rate i is fair with the ratings he/she provides. For raters motivated by content quality, $o_i = 1$, and for raters completely driven by their social network, $o_i = 0$. We make the simplifying assumption that all of a rater's selections are driven by only one of the two motivations. As part of the reputation estimation algorithm, we estimate the probability that $o_i = 1$ for each user i. To estimate o_i, we model a rater's behavior as follows: the number of ratings q_j each user provides is drawn from a distribution (this distribution need not be modeled as part of the final algorithm). Each user also has a hidden variable o_i associated with him/her. Following this, for q_j timesteps, depending on the value of o_i, the user i draws values from one of two distributions: the quality distribution (if $o_i = 1$) and his/her personal social affinity distribution (if $o_i = 0$.) Let O be a diagonal matrix where O_{ii} is the estimated objectivity value of user i.

The quality distribution ρ is defined as follows: the user selects a user at random, with the probability of user j being selected proportional to their reputation r_j. Essentially ρ is the same as the vector r, normalized. The social affinity distribution σ_i for user i is defined as the user's social network, with all members equally likely; people who are not member are assigned a small prior, to assure nonzero likelihood. We use another prior: the prior probability of selecting from the social affinity distribution defined for each user, which is the number of times the user selected a poster from his/her social network, based on historical data. We refer to this as the affinity prior π. Then given a set of selections, the posterior probability of selecting from either of the two distributions can be calculated. The quality distribution depends on O, as only users who are objective should be considered while calculating ρ. However, re-estimating ρ changes the objectivity values O for all users. We use an iterative expectation maximization based algorithm, where user objectivity and the quality distribution are alternatively estimated.

```
Step 1: *Initialization*
a) Initialize $\boldsymbol{\pi}_i$, the affinity prior probability for each user $i$.
b)Set $\rho = (I - OQP)^{-T}\boldsymbol{z}$, where $Q$ and $\boldsymbol{z}$ are initialized as described in Section III.
c) Repeat Step 2 to 4.

Step 2: *Objectivity Estimation*
For each rater $i$ in the dataset, and their ratings $\boldsymbol{s_i} = r_{ij}$.
estimate $o_i = \frac{(1-\pi_i)P(\boldsymbol{s_i}|\rho)}{(1-\pi_i)P(\boldsymbol{s_i}|\rho)+\pi_i P(\boldsymbol{s_i}|\sigma_i)}$.

Step 3: *Likelihood Estimation*
a) Calculate
$LL^{(j)} = \sum_{i=1}^{N}(1 - o_i)\log P(\boldsymbol{s_i}|\sigma_i) + o_i \log P(\boldsymbol{s_i}|\rho)$,
where $j$ is the current iteration number.
b) If $LL^{(j)} < LL^{(j-1)}$, exit.

Step 4: *Reputation Estimation*
c) Set $\rho = (I - OQP)^{-T}\boldsymbol{z}$.
```

Fig. 1. Iterative Algorithm for Reputation Estimation

6 Experimental Support

We ran experiments testing the trust algorithm on data from the user generated content website Digg[2]. Along with Reddit[3], Digg is currently one of the most used content aggregators. Digg maintains a rich, active user community and contains the necessary components for trust estimation in a content-oriented social network including: user generated content, a voting and aggregation system, and a mechanism to link users into a social network. Digg social network and endorsement data was obtained with permission from Lerman *et al.* [11]. The dataset represents one month of front page activity in 2009. For each user submitted link (story) that made it to the front page we have access to the identity of the story poster, and the identity each user who 'diggs' the link. Additionally, for each of these users we have access to the single-directional link data, indicating that a user is 'following' another, thus forming a social network. Each user has access to the activity of the users he/she follows, so that when a

[2] www.digg.com

[3] www.reddit.com

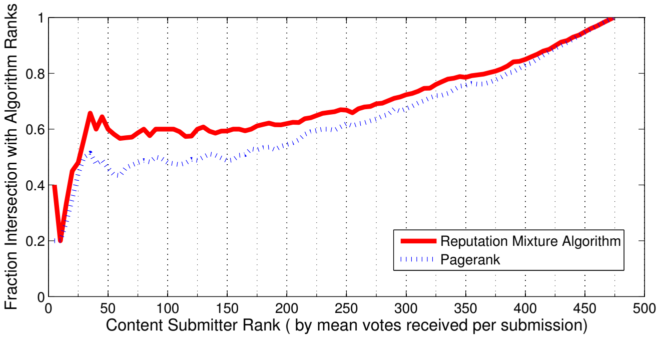


Fig. 2. Fraction of User Ranks Predicted Correctly by Reputation algorithm and Pagerank

user diggs a link, all users who follow him/her are able to see this information. A significant portion of the votes on Digg come from this process, where users find content which their friends have endorsed, a process described as a 'cascade effect' [11]. These endorsements are driven by a mixture of two classes of motivators: similarity-based and social influence-based. Similarity-based motivation occurs when a user follows a content creator because of a preference for content by that content creator, whereas social influence-based motivation occurs when a user endorses content from a creator because of a social relationship with that creator. Because these motivations are mixed, it is difficult to identify users who submit preferred content from those who are merely socially influential.

The aim of the experiments is to test whether a mixture model based approach that attempts to model social interaction dynamics can identify users relying unfairly on their social network influence to boost their reputation. This is compared to a pagerank [2] based approach that does not take into account any information about possible social motivations of voter endorsements (diggs). We expect our algorithm to identify users who provide better quality content. As a measure of content quality, we use the mean number of votes received by a user once their story is promoted to the front page, as a large majority of votes for a front-page story come from the website's broader audience, making it difficult to rely on social affiliations. For the experiments, we analyze, for each content creator/poster, the voting data for each story they have posted until it receives 30 votes. This information is used to calculate the reputation of each user using our mixture-model based algorithm. We then calculate the correlation of the reputation scores observed with the mean number of votes received per story for each poster, and compare this value to a naïve pagerank based approach.

Table 1 shows the correlation coefficient values of the reputation and pagerank scores of each story submitter with the total votes received by his/her stories. The correlation is high in both cases, but higher for the reputation algorithm. Table 2 compares the averaged reputation and pagerank scores (obtained by dividing reputation/pagerank scores with number of submissions) with the mean

Table 1. Correlation: Reputation and Pagerank scores vs submitter total votes

	Correlation Coefficient
Reputation Mixture Model	0.895
Pagerank	0.809

Table 2. Correlation: Averaged Reputation and Pagerank scores vs submitter mean votes per post

	Correlation Coefficient
Reputation Mixture Model	0.591
Pagerank	0.484

votes received per submission. We believe this to be a better measure of a content creator's quality than the aggregate number of votes, as a user can be really inconsistent in quality but still receive a large number of votes in total if he/she submits a large number of stories. However, in this case, the correlation is weaker. But the reputation algorithm still outperforms pagerank in terms of correlation.

To compare how well the two algorithms rank users by quality, we sorted scores provided by each of them in descending order, and compared that to a ranking of posters by mean number of votes received. The comparison is shown in Figure (2). The y-axis of the graph shows the fraction of users in common between the ranking of users by mean vote per submission, and the ranking generated by the algorithm. The reputation algorithm identified two of the top five ranked contributors, while the pagerank algorithm could not identify any. However, both algorithms could identify only two of the top ten. This is responsible for the initial drop in performance of the reputation algorithm from a peak. Following this the reputation algorithm consistently outperforms pagerank.

7 Conclusions

Content oriented social networks populated with user generated content are growing in popularity and diversity. Many such networks rely on large numbers of users who voluntarily generate content. This content draws in other users and creates value for the site. It is important to be able to identify the most valuable users and establish a level of trust in these users. This trust can be harnessed in the form of reputation, which is a signal that can be shared with others to drive decision making. Additionally, knowing the valuable members of a community is useful for the system designers because the designers can then implement strategies and incentive mechanisms to draw more trustworthy users to a site.

Social effects often hinder the performance of existing reputation mechanisms in UGC communities. This work presents an algorithm and demonstrates the performance of modeling user reputation in a COSN which is not biased by these social effects. We have demonstrated the performance of the algorithm on UGC data from Digg, and the results are applicable to any content-oriented social network relying on user generated content.

References

1. Barber, K., Kim, J.: Belief revision process based on trust: Agents evaluating reputation of information sources. Trust in Cyber-societies pp. 73–82 (2001)
2. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1-7), 107–117 (Apr 1998), http://linkinghub.elsevier.com/retrieve/pii/S016975529800110X
3. Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S.: Feedback effects between similarity and social influence in online communities. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 160–168. KDD '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1401890.1401914
4. Ghosh, R., Lerman, K.: Predicting influential users in online social networks. In: Proceedings of KDD workshop on Social Network Analysis (SNA-KDD) (July 2010)
5. Guo, L., Tan, E., Chen, S., Zhang, X., Zhao, Y.E.: Analyzing patterns of user content generation in online social networks. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09. p. 369. ACM Press, New York, New York, USA (2009), http://portal.acm.org/citation.cfm?doid=1557019.1557064
6. Haveliwala, T., Kamvar, S., Jeh, G.: An analytical comparison of approaches to personalizing pagerank. Technical Report 2003-35, Stanford InfoLab (June 2003), http://ilpubs.stanford.edu:8090/596/
7. Jonker, C., Treur, J.: Formal analysis of models for the dynamics of trust based on experiences. Multi-Agent System Engineering pp. 221–231 (1999)
8. Kamvar, S.D., Schlosser, M.T., Garcia-Molina, H.: The eigentrust algorithm for reputation management in p2p networks. In: Proceedings of the 12th international conference on World Wide Web. pp. 640–651. WWW '03, ACM, New York, NY, USA (2003), http://doi.acm.org/10.1145/775152.775242
9. Kemeny, J., Snell, J.: Finite Markov chains. University series in undergraduate mathematics, VanNostrand, New York, repr edn. (1969), http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+232428700&sourceid=fbw_bibsonomy
10. Lee, J., Antoniadis, P., Salamatian, K.: Faving Reciprocity in Content Sharing Communities: A Comparative Analysis of Flickr and Twitter. In: Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on. pp. 136–143. IEEE (2010), http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5562780
11. Lerman, K., Ghosh, R.: Information contagion: an empirical study of spread of news on digg and twitter social networks. In: Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM) (May 2010)
12. Lerman, K., Jones, L.: Social browsing on flickr. In: Proceedings of 1st International Conference on Weblogs and Social Media (ICWSM-07) (2007)
13. Lussier, J., Raeder, T., Chawla, N.: User Generated Content Consumption and Social Networking in Knowledge-Sharing OSNs. Advances in Social Computing pp. 228–237 (2010), http://www.springerlink.com/index/F6345441592572X6.pdf
14. Pavlov, E., Rosenschein, J., Topol, Z.: Supporting privacy in decentralized additive reputation systems. Trust Management pp. 108–119 (2004)
15. Resnick, P., Kuwabara, K., Zeckhauser, R., Friedman, E.: Reputation systems. Communications of the ACM 43(12), 45–48 (2000)

16. Rose, K.: Digg v4: release, iterate, repeat. (Jul 2010), http://kevinrose.com/blogg/2010/8/27/digg-v4-release-iterate-repeat.html
17. Sabater, J., Sierra, C.: Reputation and social network analysis in multi-agent systems. In: Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1. pp. 475–482. ACM (2002)
18. Thurman, N.: Forums for citizen journalists? Adoption of user generated content initiatives by online news media. New Media & Society 10(1), 139 (2008)
19. Van Alstyne, M., Brynjolfsson, E.: Global Village or Cyber-Balkans? Modeling and Measuring the Integration of Electronic Communities. Management Science 51(6), 851–868 (Jun 2005), http://mansci.journal.informs.org/cgi/doi/10.1287/mnsc.1050.0363
20. Weng, J., Lim, E., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 261–270. ACM (2010), http://portal.acm.org/citation.cfm?id=1718520
21. Yolum, P., Singh, M.: Engineering self-organizing referral networks for trustworthy service selection. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans 35(3), 396–407 (2005)

Sources of Stereotypical Trust in Multi-Agent Systems

Chris Burnett[1], Timothy J. Norman[1], and Katia Sycara[1,2]

[1] University of Aberdeen
{cburnett@abdn.ac.uk, t.j.norman}@abdn.ac.uk
[2] Carnegie Mellon University
katia@cs.cmu.edu

Abstract. In highly dynamic and open multi agent systems, where agents must interact with each other when pursuing their own goals, success may depend on the abilities of agents to appropriately judge the trustworthiness of their potential partners. However, where the rate of agent turnover is high, it can be difficult to obtain sufficient evidence from which to build trust. To address this, recent approaches have proposed a form of *stereotypical* trust, using visible *features* of agents to form generalised trust assessments. However, these approaches generally assume that features are explicitly observable, and are not concerned with what form these features take, or how they may be derived. In this paper, we argue that knowledge about social relationships between agents may provide a useful source of feature information. We discuss a number of sources of features, and outline potential strategies for improving the effectiveness of stereotypical trust evaluation mechanisms. Finally, we present trusted group formation as a potential application domain for stereotyping techniques, and discuss key areas for future work in this direction.

Keywords: ad-hoc teams, reputation, stereotypes, trust

1 Introduction

In highly dynamic and open multi-agent systems, agents may be deceptive or exhibit varying degrees of competence in their stated capabilities. When agents must interact with others to achieve their own goals, their success becomes dependant on the motivations and behaviours of other agents in the society. Therefore, there have been efforts to enable agents to rapidly build models of 'trust' within their environment, with the aim of helping agents to learn about the trustworthiness of others and subsequently identify the most trustworthy partners for interaction [22, 19, 14].

This problem is made more difficult when the population of a multi-agent system is highly dynamic, i.e. when agents frequently join and leave, or when the structure of the society is dynamic, i.e. when agents have large numbers of competences, or may often change roles or social positions within the society. Under these circumstances, it can be difficult for an agent to obtain sufficient evidence about a potential partner's past behaviour in order to build predictive models of trust capable of making evaluations about new, unknown agents. In the worst case, where the agent population is highly fluid, agents may never have be capable of forming trust evaluations using traditional direct- or reputation-based mechanisms.

The notion of stereotypical or category-based trust has been recently proposed as a way of addressing these issues [3, 7]. In real-world scenarios, people may be ascribed certain *features* which correlate with trustworthiness. For example, employees of a reputable firm may be more likely to be trustworthy in a particular task than employees of a disreputable firm. Over time, through interacting with individuals, people learn to generalise their trust to trust in sets of features. This *stereotyping* process plays a vital role in reducing the initial uncertainty in human interactions [18]. However, these underlying correlations may also exist in artificial environments. Falcone and Castelfranchi [7] discuss the generalisation of trust within the context of cognitive trust models [4], and provide a formal characterisation of how an agent's beliefs may be generalised based on the perceived featural similarity of other agents and tasks.

In this paper, we aim to address an important question associated with the use of such mechanisms: what kinds of social or contextual knowledge can be exploited in order to build these stereotypes? Existing approaches consider stereotypes which are activated by visible, abstract features, such as nationality or gender. However, agents may be situated within various complex (and possibly overlapping) social contexts. These social contexts may affect the trustworthiness or reliability of agents when interacting with others from a different social context. Therefore, it is important to consider these contexts when forming stereotypes. For example, agents may be considered untrustworthy (or trustworthy) because of the relationships they maintain with other individuals or types.

This reflects the intuition behind common proverbial expressions such as “tell me who your friends are, and I'll tell you who you are”, and “birds of a feather flock together”. We may, for example, expect that an agent who maintains the relationship *friend* with a large number of convicted criminals might not be very trustworthy. However, this example shows that it is important to also consider the type of relationship involved: if an agent maintains a relationship ‘counsellor’ with a large number of such criminals, then that agent is likely acting in a professional capacity, and its trustworthiness should not be affected. While this reasoning may seem biased, we argue that if such relationships can can effect trustworthiness, then they should be considered when forming stereotypes.

The rest of the paper proceeds as follows. In Section 2 we outline existing mechanisms for forming stereotypical trust. In Section 3 we discuss a number of implicit sources of featural evidence which could be used by stereotypical trust mechanisms. In Section 4, we present the problem of trusted group formation as a potential application of these feature sources. Finally, we discuss avenues for future work and conclude in Sections 5 and 6.

2 Stereotypical Trust

In existing stereotypical trust approaches, agents learn by interacting with others, and observing both the interaction outcomes, and the visible attributes (or *features*) of partners. In this paper we are concerned primarily with the potential *sources* of information from which stereotypes may be formed. In order to motivate our discussion, we present in this section an overview of existing techniques for forming stereotypical trust evaluations.

In [16], the authors describe *StereoTrust*, which attempts to build stereotypes on the basis of agents' observed membership of particular groups. In this model, features are derived from explicit group memberships. However, the authors do not present mechanisms for identifying groups from salient low-level features, instead assuming that instances of groups are explicitly provided. In [3], a mechanism is described which allows stereotypical groups to be identified based on observed correlations between features and behaviour. This model also allows for the sharing of stereotypical opinions, which is useful when no individuating reputational evidence is available anywhere within the society. The approach of Hermoso et al. [11] proposes a centralised mechanism which assigns new features (corresponding to *roles*) to agents according to globally observed behavioural trends. In contrast to the previously described approaches, this mechanism attempts to *create* stereotypical relationships in order to assist agents in locating suitable partners.

In these works, the use of stereotypical trust evaluations in highly dynamic settings is shown to increase the average utility gain of trustors. Generally, the stereotyping model of an agent a can be represented as a function S_a which, given a set of features from a global feature set $\mathcal{F}$, produces an *a priori* trust evaluation for any trustee possessing those features:

$$S_a : 2^{\mathcal{F}} \rightarrow \mathbb{R} \tag{1}$$

Here, we assume the output of the function is a real-valued trust estimate which may be used when direct or reputational evidence is unavailable, or combined with existing trust estimates as a bias. The aim of a stereotypical trust approach is to allow agents to learn these functions from their history of interactions within a society. Once a trustor has gathered enough evidence to build a stereotyping model, the resulting stereotypical evaluations may be used when other forms of evidence are unavailable.

Decision trees [2] provide an appropriate model for capturing the behaviour of stereotyping functions. By representing the stereotyping function in this way, we can make use of well-known techniques for inducing decision trees from labelled examples [8, 15]. Furthermore, it is possible to encapsulate all of an agent's stereotypes about others regarding features in $\mathcal{F}$ in one concise structure. Each node of the tree represents a particular feature, and branches from nodes are followed depending on the perceived value of the feature represented by that node. Each leaf of the tree represents the stereotypical base rate (or a function producing a base rate) that will be applied to all classification examples reaching that leaf. Figure 2 shows an example of a simple decision tree being used to classify an agent with a visible feature vector $F = \{a, \neg b, \neg c, d\}$. The resulting path through the tree results in a predicted stereotypical evaluation for the agent, based on the feature vector F. When evaluating an agent y for which we have no evidence, the stereotype tree can be used to obtain an estimated *a priori* trust value.

Each path through the tree then represents a particular rule, or *stereotype*, which produces in an *a priori* trust evaluation for a given feature input. In the remainder of this paper, we will focus on potential information sources from which features may be derived, providing additional evidence for stereotypical trust mechanisms.

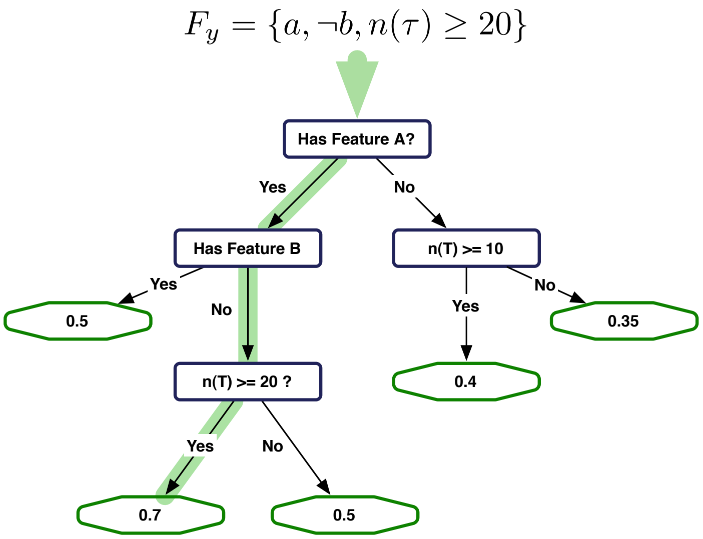


Fig. 1. Example decision tree encoding a stereotyping function

3 Feature Sources

Until now, we have discussed members of $\mathcal{F}$ in an abstract way. We have assumed that they are directly visible features that an agent displays to the rest of the society. However, agents may not necessarily display their predictive features in this way. However, there may be many other *implicit* features which can be derived from publicly observable information about an agent, which may correlate with trustworthiness. In the following sections, we will discuss some sources of such *featural* evidence.

3.1 Social Networks

Interactions in multi-agent systems may conceivably take place within the context of a *social network*. Agents might maintain explicit relationships with each other, which have some significance within the society. Examples of social networks include family ties, organisational hierarchies, and trust networks [13, 10], with relationships such as “a is a brother of b”, “a is superior to b”, and “a trusts b”, respectively. Social networks, as a means of representing and reasoning about social relationships, were pioneered in sociological domains for the analysis of communities, organisations, and political structures [23]. However, they are also useful for the representation of abstract structures, and have received significant attention in multi-agent systems as well, having been used to represent other concepts, such as influence [17], dependance [25], trust [5] and reputation [21]. In this section, we will discuss the applicability of such networks as a source of feature information for stereotypical trust approaches.

Social networks can be represented as directed graphs, where nodes represent entities in a structure, and arcs represent some type of relationship. Edges between nodes can be labelled with the attributes of the relationship which are of interest. There may be *implicit* behavioural effects resulting from these relationships. For example, consider a hierarchical-type relationship between two agents, whereby one is 'superior' to the other, such as "a is the manager of b". The agent b may be normatively compelled to behave in a respectful way towards the manager a, but not towards other agents with whom no such relationship exists. Essentially, agent b's loyalties lie with a. Even this simple example has implications for trust. For example, a third agent c, who has no explicit relationship with b, may be able to delegate tasks to b. However, they may be interrupted by a's requests, which take priority. From the perspective of c, b's trustworthiness may be decreased by the existence of the relationship with a, because c knows that he can never fully align b's interests with his own.

A similar example of implicit consequences arising from social relationships is that of collusion. If two agents are related in a way which implies their interests may be closely aligned (such as friendship, co-worker, or family ties for example), then there may be a motivation for these agents to collude with each other when interacting in the society, to the detriment of others.

When such relationships (and their associated implicit effects) exist within a society, a stereotyping approach may be capable of detecting them. We propose viewing relationships between agents as features. For example, the feature representing a relationship of type R from one agent, a, to another, b, can be represented simply by a possessing the feature Rb, and b possessing the feature aR. Bi-directional relationships can be represented as two features, one for each direction of the relationship.

For example, consider the social network illustrated in Figure 2. Two types of relationships are present here, labelled A (dashed lines) and B (solid lines). In this example, the agent d could be annotated with a feature vector $F_d = \{Bc, Ba, Be, Bb, Ag, gA\}$. This representation can be easily transformed into an adjacency matrix, which can be used directly by our learning mechanism.

To illustrate our example further, consider that the relationship type B represents the relationship *instructs*. If an agent a instructs an agent b, this could be captured by a possessing the feature Bb, and b possessing the feature aB. Consider that the agent d in Figure 2 is a particularly effective instructor, whose students are highly likely to be competent and trustworthy. It would be difficult for trustors to learn this information simply by interacting with the students of d (circled), as there could by a number of instructors responsible for a single student. For example, agent a in Figure 2 is instructed by agents c, d and b. However, by using a stereotyping approach, trustors may be able to learn, from experience, that all the students of d are likely to be trustworthy. In terms of a stereotype, agents that possess the feature dB benefit from a higher degree of initial trust than agents without the feature, due to high occurrence of good performance among students of d. By considering relationships as features, stereotyping can help uncover behavioural correlations associated with the social structure itself, when this information is available.

It may be possible to infer the existence of significant social groups from the presence of certain relationships between the individual members. For example, in Figure 2, the agents related to agent d by the *instructs* relationship may be considered a social group, defined as "agents who were instructed by d", or "former students of d".

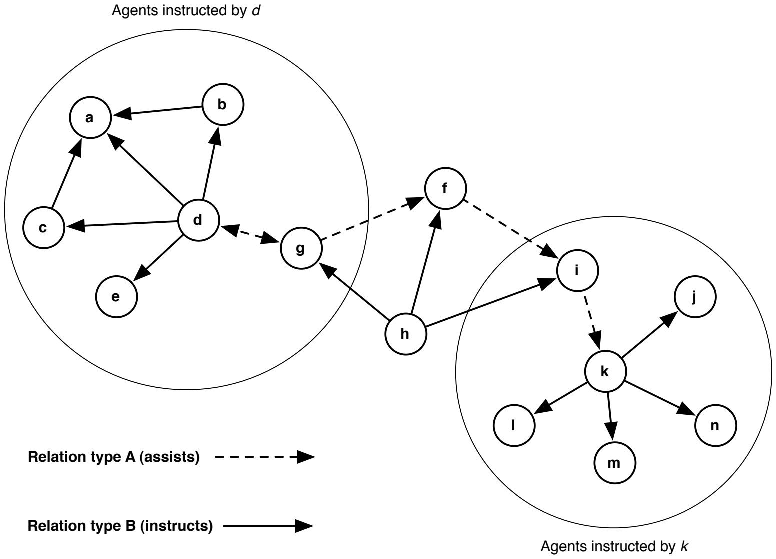


Fig. 2. Example social network with multiple relationship types.

Shared views and norms can arise from these relationships, potentially affecting the perceptions and behaviours of group members towards members of other groups. For example, agents in this group are likely to share certain practices with each other, and with d, as a result of this relationship. The presence of these loosely defined groups may therefore have an effect on the trustworthiness of different groups, from the perspective of others. Continuing our example, if we do not trust agent d because of his practices in a given activity, we may not trust former students of d either.

Popular social networking systems, such as Twitter[3] and Facebook[4], provide a wealth of such complex social relationships which may be used to construct stereotypes. For example, twitter is a service which allows users to publish short messages to a potentially large readership. Users can explicitly follow others, creating an explicit social relationship and implicitly indicating interest (and possibly trust). Users can also *mention* others, reply to message authors, create threads of interest (known as *hashtagging*) and forward received messages to their own subscribers (known as *retweeting*). These latter aspects are implemented through the use of syntactic conventions in messages. By analysing these messages, a detailed picture of a social network may be constructed which, using the process described above, may provide a useful source of feature-based evidence when forming stereotypical trust evaluations. With some additional discretisation, we can derive binary features from these activities.

For example, if an agent a mentions another agent b on a daily basis, we may assign a feature $frequently\_mentions\_b$ to a. Assume that b represents some company

[3] http:\\www.twitter.com

[4] http:\\www.facebook.com

manufacturing a product, and that this company employs a large number of agents to disseminate favourable comments in the society. Excessive mentions may not necessarily indicate deceptive behaviour; the products of b may be particularly noteworthy, for better or for worse. If, however, by interacting within the society, we form a stereotype indicating that agents with the $frequently\_mentions\_b$ feature are less trustworthy when recommending products than those without, we may assume that such agents are indeed biased.

3.2 Competency over time

An interesting application of stereotypical trust approaches may lie in addressing situations where the competency of agents may vary over time in a predictable way according to experience. Recent work on trust models, such as the probabilistic model proposed by [26], focuses on maintaining high performance when the behaviour of trustees can rapidly, adversely and randomly change. In such cases, it is important that the trust model respond quickly. However, in many domains, the behaviour of agents may change in a predictable manner over time. For example, agents that perform recognition or identification tasks will become more reliable as they build up a corpus of experiences. New agents will initially perform poorly, but will become steadily more competent the more they are interacted with. The way in which this happens may vary however; agents with a poor learning approach will require more interactions to achieve a level of accuracy considered sufficient by the society, whereas agents with more effective approaches will become competent faster.

In such cases, trustors should be able to form stereotypical assumptions about the trustworthiness of potential partners, based on observable attributes which indicate their level of *experience*, such as the number of times they have performed a particular task. For example, a particularly ‘slow-learning’ type of agent may require, on average, 40 interactions to reach a certain level of competence, while another ‘fast-learning’ type of agent may require only 20.

These relationships can be addressed by a stereotypical trust approach. Information about an agent’s accumulated experience in different tasks can be converted into features. For example, we may create a feature token for every ten instances of a task performed by an agent, creating training ‘milestones’. If an agent has performed a task τ 23 times, we may signify this with the feature ‘$n(\tau) \geq 20$’, meaning “performed τ at least 20 times”. Alternatively, we can allow the splitting function of the stereotype learning mechanism to determine a suitable discretisation. By interacting with agents with different levels of experience, trustors can build stereotyping models based on these experiential features, which can then be used to predict the trustworthiness of new, unknown agents. These models then represent ‘learning curves’ for tasks, as they estimate the trustworthiness of agents given the experience they have accumulated

As features here are simply symbols, experiential features can be used alongside explicit features. For example, agents with certain features may learn faster than others. Figure 3 provides an example stereotype including task experience features.

A key disadvantage to this approach is the requirement for information about a trustee’s past experience in a task. We have assumed here that this would be provided by the trustee, but this introduces the potential for the trustee to lie about its experience.

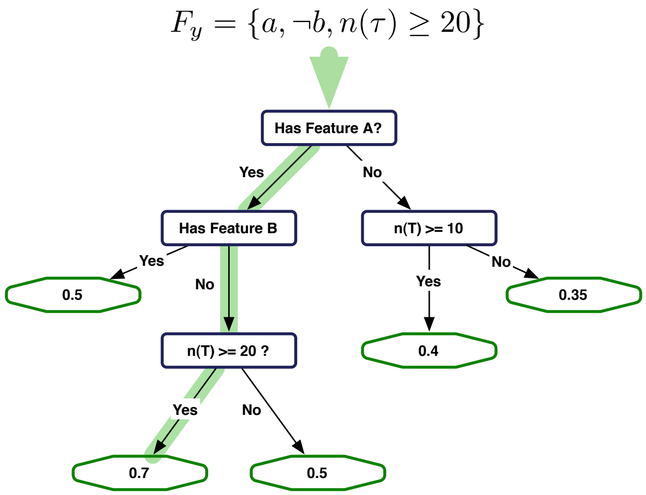


Fig. 3. Example stereotype model with features representing trustees' accumulated experience.

One way of dealing with such a problem may be to employ a 'certified reputation' approach, as advocated by Hyunh et al. [12], whereby trustors agree to provide 'certified' ratings to their trustees, which can then be provided to future partners. In such an approach, the authors require that certified testimonies of trustors are digitally signed to ensure authenticity.

However, using stereotyping models in this way may lead naturally to an undesirable and paradoxical result; as stereotypes about experience become established, trustees entering the society with accumulated experience from outside the society will begin to be preferred over those without. Incoming trustees with less experience will gradually become less likely to be selected for interaction than those already possessing some experience. As a result, inexperienced trustees may be precluded them from gaining the necessary experience[5], essentially 'freezing out' less experienced trustees. While this situation is favourable to the trustors, it may be harmful in the long run, as it prevents the efficient distribution of interaction within the society.

3.3 Interaction Stereotypes

Until now, we have assumed that stereotypes represent the observations of one agent about other learned 'types' in the society, based on the observable features of those types. However, it is equally possible that differences in trustworthiness arise not only from features of the trustee, but of either agent in a dyadic interaction. It is possible, for example, that agents with certain features are positively or negatively *biased* towards agents with certain other features. It is therefore important to consider stereotypes which capture salient features of both agents.

[5] This type of situation may be familiar to many new university graduates upon entering the job market.

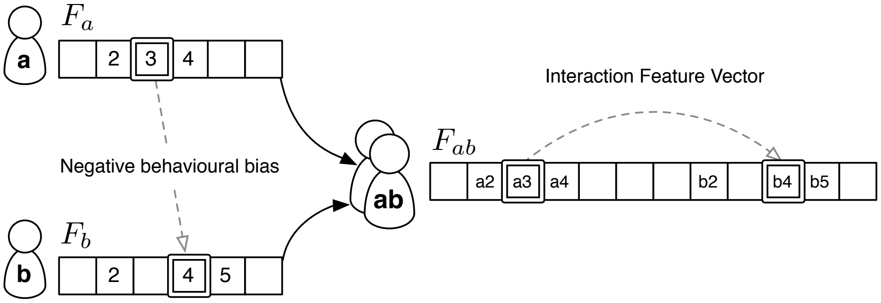


Fig. 4. Converting an observation of dyadic behaviour into a stereotypical interaction observation.

To address this problem, we introduce the notion of *interaction* stereotypes. These are stereotypes which apply to pairs of agents, rather than single individuals. For example, the statement “trustors with the feature j perform well in interactions with trustees with the feature k” is an informal example of such a stereotype.

We can develop these interaction stereotypes in the following way. After an agent has observed an interaction between any two other agents (where one party may be the agent itself), the feature vectors of both participants are concatenated to form an *interaction feature vector* for the dyad. A symbol a or b is appended to each feature symbol to indicate whether the feature’s owner was playing the role of a trustor or trustee. The observed outcome of the interaction is then associated with the interaction feature vector to create a training instance for a stereotypical trust model.

Figure 4 illustrates this process. A hidden behavioural bias exists between agents with feature 3 and those with feature 4. As a result, agents with the feature 3 are likely to be less trustworthy than normal when interacting with agents with feature 4 (we refer to this as negative behavioural bias).

These examples can now be used to train a stereotyping model to detect feature-behavioural correlations between pairs of agents. From a number of observations of different agents with features 3 and 4 interacting, we may form a stereotype which predicts a low likelihood of a good outcome in any interaction where the interaction feature vector contains both features 3 and 4. In our previous example, we would expect this negative stereotype to be activated when observing an interaction involving the features $a3$ and $b4$.

By gathering reputational evidence from other agents in the society, and using this to train a *global* stereotype model, an agent can build a picture of the network of stereotypical relationships between other types. Figure 5 shows an example of such a network. Stereotypical relationships are represented by arcs labelled positively or negatively, depending on the modality of the relationship. In this example, agents of type X and Z share a positive stereotypical opinion about each other. Agents of type X are also stereotypically positive about agents of type Y, but this is not reciprocated. All types have a positive stereotypical opinion about type Y.

Note that the network does not contain a bi-directional arc between each agent type. This is because we might not detect a significant feature-behavioural correlation be-

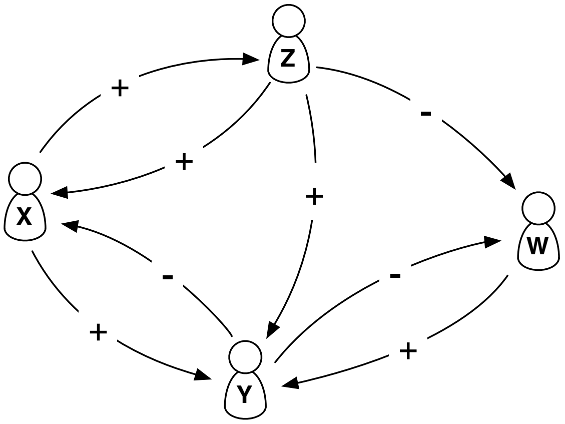


Fig. 5. An example stereotypical trust network.

tween each type. Also, note that the types W, X, Y and Z are products of the stereotyping mechanism; they need not be explicit groups. They will be defined simply as sharing a number of common stereotypical features.

4 Group Formation

An interesting example application of the above approaches may arise when it is desirable to explicitly form groups, teams or coalitions of agents who are 'most trusted' for their assigned roles, from the perspective of some agent responsible for forming the group (whom we will refer to as the group *owner*).

We assume that the owner has some knowledge about the desired group structure G, expressed as a directed graph, where nodes represent *roles* that agents can play, and edges represent relationships between those roles. Figure 6 shows an example of a simple group structure. When the group enacts its task, agents playing roles connected by an edge will decide whether to delegate the task indicated by the edge label. A particular *instantiation* of a group structure is an assignment of agents to roles, such that each role is played by exactly one agent when the group begins to interact. In this example, we do not consider time as a factor; once the group task is delegated to a particular instantiation, we assume all agents interact at once, and the results of those interactions are returned to the group owner.

Figure 6 shows a number of candidates, numbered r_1 to r_8, assigned to group roles. Consider that these agents have no prior experience of working together, as may often be the case, and that stereotypical biases exist between agents; odd-numbered agents are biased against even-numbered agents, and vice versa. Both odd- and even-numbered agents place higher degrees of initial trust in others of the same type. The problem for the owner is to find the best set of agents from global society, and the best assignment of trustees to roles, so that the initial conditions of trust within the team are maximised.

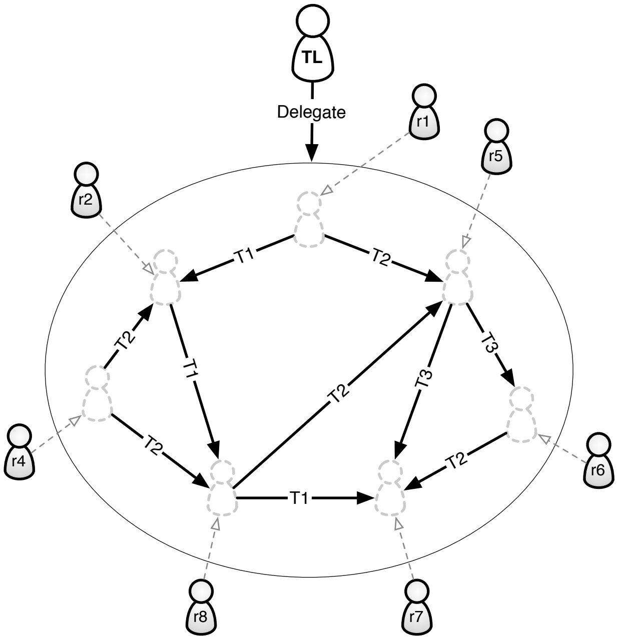


Fig. 6. A simple group structure, with instantiation candidates.

Ideally, agents would be assigned to roles such that behavioural biases increase the likelihood of a good group outcome. Already, this problem can be formulated as a combinatorial optimisation problem, and a number of techniques exist to enable the formation of optimal groups or coalitions in multi-agent systems [20, 24, 9].

In particular [24] discusses the use of previously identified agent *types* in simplifying the problem of forming optimal coalitions of agents. Stereotyping approaches can help to identify these types based on the relationships between agent behaviours and sets of features, and subsequently simplify the problem of identifying possible groups. For example, when a number of agents match an observing agent's interaction stereotype, they will appear stereotypically identical. Therefore, in simplifying the group formation problem, stereotypically identical agents can be considered as a single virtual agent representing the stereotype, with the group owner being indifferent between those agents.

When groups are explicitly formed in this way, the group owner may himself become a useful feature for stereotyping. This is explicitly considered as a source of trust evidence by [18]:

> "Each member assumes that the contractor has either had the requisite experience with others, or, at the very least, that he or she has 'asked around' and 'checked them out'. Thus, trust in the contractor's presumed care in composing the temporary group serves as a proxy for individual knowledge or experience with others' reliability of competence." [18]

Intuitively, if a particular group owner is highly competent in assembling groups in which all members are highly satisfied most of the time, then group members may begin to trust the judgement of the group owner. This, in turn, may lead to an increased initial degree of trust between unfamiliar group members, i.e. when they are unable to form opinions from direct experience, or from the experiences of their peers.

This effect could be modelled by employing stereotyping techniques to develop trust in the group owner. When a group is formed by an agent z, each member agent is labelled with a new feature, such as go^z, which indicates that the agent was chosen by the group owner. As agents interact within the team, and build stereotypes, their behaviours will reflect on this new feature go^z. For example, if z always chooses competent agents for groups, then the feature go^z will correlate highly with trustworthiness in the society. In subsequent groups formed by z, possession of the feature go^z will then result in a higher initial level of trust within the group, due to the influence of z.

5 Future Work

While we have described a number of sources of stereotypical features in this paper, our discussion has remained abstract. An interesting avenue for future work involves the evaluation of these techniques within real multi-agent systems, whether they be populated by artificial and autonomous agents, by agents representing real human personae (such as e-commerce or social networking websites) or a mixture of both. It remains to be seen whether using these feature sources can aid in the detection of such subtle stereotypical behavioural variations in real-world systems. It is therefore desirable to evaluate these approaches using datasets from real social network or e-commerce systems, where feature-behavioural correlations are not artificially controlled.

The problem of effective team and coalition formation presents a potential application area for stereotypical trust models. Efforts have been made to understand how team and coalition formation mechanisms may be extended to consider notions of trust [1, 6, 9]. However, to the best of our knowledge, the application of stereotypical trust to the problem of group/team formation in highly dynamic environments has yet to be addressed. Besides allowing unknown candidates to be initially classified, stereotyping techniques may provide an initial benefit by helping to reduce the search space of team formation algorithms, by reducing the space of available candidates to a smaller space of candidate 'types'.

One drawback of our current approach is that the generation of simple features from real-valued observations still requires some custom (and subjective) discretisation on behalf of the observing agent. This necessarily affects the subsequent stereotypical learning process. This problem may be addressed through the use of learning techniques which permit the use of real-valued 'features' to be used directly. Alternatively, it may be beneficial to employ automatic discretisation based on clustering techniques.

Another key future direction involves exploiting ontological relationships between the features agents possess. In this work, we have considered stereotypical features which are 'flat', in that all features are assumed to be independent of one other. However, in practical applications, there may be sets of features for which hierarchical relationships exist. That is, there may exist sub-features which are instances of other super-features. If such relationships exist, they could be exploited when forming stereotypes.

For example, the features *cardiologist*, *general practitioner*, *optician* and *surgeon* could all be considered sub-features of a more general feature *doctor*. In this example, it is reasonable to presume that feature-behaviour correlations could exist for agents possessing the *doctor* feature. For instance, we would expect that all doctors, regardless of their chosen specialisation, can be expected to be competent at administering first-aid.

In certain cases, there may be benefits to considering these relationships in the stereotyping process. Firstly, in highly dynamic environments, where agents are described by a large number of features, and the number of possible features agents can possess is large, we may arrive at a similar problem when forming stereotypes as we face when forming trust with no stereotyping model. That is, agents may not encounter known *features* frequently enough to form useful stereotypes.

Extending our medical example above, assume that we have no knowledge of the ontological relationship between sub-types of *doctor*. Furthermore, the general feature *doctor* is not publicly visible. If we observe each of the possible doctor sub-types once in a first-aid task, we have significant evidence about the trustworthiness of agents possessing features which are a sub-type of *doctor*. However, without knowledge of this ontological relationship, we are not able to make this generalisation. Therefore, we may only be able to identify a weak feature-behavioural correlation between each of the specific doctor types individually. Future work will investigate the use of ontological relationships between features to improve the stereotyping process with larger, richer feature sets.

6 Conclusion

In highly dynamic multi-agent systems, it is important to consider various sources of available evidence when evaluating the trustworthiness of others. By forming stereotypes, agents can generalise their partners to *types* which are more resistant to the degree of agent turnover within the society. These generalisations attempt to capture the relationships between *features* of agents and trustworthiness. Since these relationships may be of a subtle nature, it is important to identify the sources of features which enable their discovery.

In this paper, we have discussed potential sources of feature information that may be derived from analysis of the characteristics of observable social relationships of potential partners. Even when agents frequently join and leave the society, familiar patterns may emerge which allow effective stereotypes to form. When the behaviour of agents can be affected by subtle social relationships, or when agents intentionally attempt to conceal their explicit features, much may be learned by incorporating these forms of featural evidence when forming stereotypes.

Acknowledgements

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

1. Barber, K.S., Ahn, J., Budalakoti, S., DeAngelis, D., Fullam, K.K., Jones, C.L.D., Sui, X.: Agent trust evaluation and team formation in heterogeneous organizations. In: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems. pp. 261:1–261:2. AAMAS '07, ACM, New York, NY, USA (2007), `http://doi.acm.org/10.1145/1329125.1329442`
2. Breiman, L.: Classification and regression trees. Chapman & Hall (1984)
3. Burnett, C., Norman, T.J., Sycara, K.: Bootstrapping trust evaluations through stereotypes. In: Proceedings. of 9th International Conference on Autonomous Agents and Multiagent Systems. pp. 241–248 (2010)
4. Castelfranchi, C., Falcone, R.: Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In: Proceedings of the 3rd International Conference on Multi Agent Systems. pp. 72–79 (1998)
5. Castelfranchi, C., Falcone, R., Marzo, F.: Being trusted in a social network: Trust as relational capital. Trust Management pp. 19–32 (2006)
6. Chalkiadakis, G., Boutilier, C.: Bayesian reinforcement learning for coalition formation under uncertainty. In: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems. vol. 3, pp. 1090–1097 (2004)
7. Falcone, R., Castelfranchi, C.: Generalizing trust: Inferencing trustworthiness from categories. In: Lecture Notes in Computer Science. vol. 5396, pp. 65–80 (2008)
8. Frank, E., Wang, Y., Inglis, S., Holmes, G., Witten, I.: Using model trees for classification. Machine Learning 32(1), 63–76 (1998)
9. Griffiths, N., Luck, M.: Coalition formation through motivation and trust. In: Proceedings of the second international joint conference on Autonomous agents and multiagent systems. pp. 17–24. ACM New York, NY, USA (2003)
10. Hang, C.W., Wang, Y., Singh, M.: Operators for propagating trust and their evaluation in social networks. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2. pp. 1025–1032. International Foundation for Autonomous Agents and Multiagent Systems (2009)
11. Hermoso, R., Billhardt, H., Ossowski, S.: Role Evolution in Open Multi-Agent Systems as an Information Source for Trust. In: Proceedings. of 9th International Conference on Autonomous Agents and Multiagent Systems (2010)
12. Huynh, T., Jennings, N., Shadbolt, N.: Certified reputation: how an agent can trust a stranger. In: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems. pp. 1224–1232. ACM (2006)

13. Jøsang, A., Hayward, R., Pope, S.: Trust network analysis with subjective logic. In: Proceedings of the 29th Australasian Computer Science Conference-Volume 48. pp. 85–94. Australian Computer Society, Inc. Darlinghurst, Australia, Australia (2006)
14. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decision Support Systems 43(2), 618–644 (2007)
15. Kalles, D., Morris, T.: Efficient incremental induction of decision trees. Machine Learning 24(3), 231–242 (1996)
16. Liu, X., Datta, A., Rzadca, K., Lim, E.: Stereotrust: a group based personalized trust model. In: Proceeding of the 18th ACM conference on Information and knowledge management. pp. 7–16. ACM (2009)
17. McCallum, M., Vasconcelos, W., Norman, T.: Organisational change through influence. Autonomous Agents and Multi-Agent Systems 17(2), 157–189 (2008)
18. Meyerson, D., Weick, K., Kramer, R.: Swift trust and temporary groups. In: Kramer, R., Tyler, T. (eds.) Trust in Organizations: Frontiers of Theory and Research, pp. 415–445. Sage Publications Inc (1996)
19. Ramchurn, S.D., Hunyh, D., Jennings, N.R.: Trust in multi-agent systems. Knowledge Engineering Review 19(01), 1–25 (2004)
20. Rathod, P., des Jardins, M.: Stable team formation among self-interested agents. In: AAAI Workshop on Forming and Maintaing Coalitions in Adaptive Multiagent Systems (2004)
21. Sabater, J., Sierra, C.: Reputation and social network analysis in multi-agent systems. In: Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1. pp. 482–490. ACM (2002)
22. Sabater, J., Sierra, C.: Review on Computational Trust and Reputation Models. Artificial Intelligence Review 24(1), 33–60 (2005)
23. Scott, J.: Social network analysis. Sociology 22(1), 109 (1988)
24. Shrot, T., Aumann, Y., Kraus, S.: On agent types in coalition formation problems. In: van der Hoek, Kaminka, Lesperance, Luck, Sen (eds.) Proceedings. of 9th International Conference on Autonomous Agents and Multiagent Systems. pp. 757–764 (2010)
25. Sichman, J., Conte, R.: Multi-agent dependence by dependence graphs. In: AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems. pp. 483–490. ACM, New York, NY, USA (2002)
26. Vogiatzis, G., MacGillivray, I., Chli, M.: A probabilistic model for trust and reputation. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). pp. 225–232 (2010)

An abstract framework for reasoning about trust

Elisabetta Erriquez, Wiebe van der Hoek, and Michael Wooldridge {e.erriquez, Wiebe.Van-Der-Hoek, mjw}@liverpool.ac.uk

Department of Computer Science, University of Liverpool, United Kingdom

Abstract. In settings where agents can be exploited, trust and reputation are key issues.In this paper, we present an abstract framework that allows agents to form coalitions with agents that they believe to be trustworthy. In contrast to many other models, we take the notion of *distrust* to be our key social concept. We use a graph theoretic model to capture the distrust relations within a society, and use this model to formulate several notions of mutually trusting coalitions. We then investigate principled techniques for how the information presented in our distrust model can be aggregated to produce individual measures of how trustworthy an agent is considered to be by a society.

Keywords: models of trust, society models, distrust

1 Introduction

The goal of coalition formation is to form robust, cohesive groups that cooperate to the mutual benefit of all the coalition members. With a small number of exceptions, existing models of coalition formation do not generally consider trust [1, 7]. In more general models [9, 6], individual agents use information about reputation and trust to rank agents according to their level of trustworthiness. Therefore, if an agent decides to form a coalition, it can select those agents it reckons to be trustworthy. Alternatively, if an agent is asked to join a coalition, it can assess its trust in the requesting agent and decide whether or not to run the risk of joining a coalition with it.

However, these models are inherently *local*: they lack a *global* view. [9] and [6] for instance consider the trust that agents that are asked to join a coalition, have in the agent initiating the coalition, but the mutual trust among the potential members is not taken into account. In this paper, we address this limitation. We propose an abstract framework through which autonomous, self-interested agents can form coalitions based on information relating to trust. In fact, we use *distrust* as the key social concept in our work. Luckily, in many societies, trust is the norm and distrust the exception, so it seems reasonable to assume that a system is provided with information of agents that distrust each other based on previous experiences, rather than on reports of trust. So, we focus on how distrust can be used as a mechanism for modelling and reasoning about the reliability of others, and, more importantly, about how to form coalitions that satisfy some stability criteria. We present several notions of mutually trusting coalitions and define different measures to aggregate the information presented in our distrust model.

Taking distrust as the basic entity in our model allows us to benefit from drawing an analogy with a popular and highly influential approach within *argumentation* *theory* [10]. Specifically, the distrust-based models that we introduce are inspired by the *abstract argumentation frameworks* proposed by Dung [3]. We show that several notions of stability and of extensions in the theory of Dung naturally carry over to a system where distrust, rather than attack, is at the core. We extend and refine some of these notions to our trust setting.

Section 2 gives the formal definition of the framework presented, while Section 3 explains how the information presented in abstract trust frameworks can be *aggregated* to provide a single measure of how trustworthy individuals within the society are. Section 4 concludes the paper and presents some possible avenues for future work.

2 A Framework for trust ... based on distrust

We assume that agents have some incentive for sharing their evaluations of the other agents in the community. Although much previous work deals with trust relationships, in our approach, we consider only *distrust* relationships between the agents. In real life, people do not always share their positive evaluation about others, but they are more inclined to report bad experiences, as a warning to other people and as a way to affect the reputation of the person the bad experience was with, as showed by the large research around *negative word of mouth* [13]. We believe that networks based on trust relations can be transformed into distrust-based networks, under some weak assumptions, but that is not the research here; we simply assume the distrust relations between agents *i* and *j* to be given, intended as agent *i* having none or little trust in agent *j*. More precisely, when saying that agent *i* distrusts agent *j* we mean that, in the context at hand, agent *i* has insufficient confidence in agent *j* to share membership with *j* in one and the same coalition.

Definition 1 *An* Abstract Trust Framework *(*ATF*), S, is a pair:* $S = \langle Ag, \leadsto \rangle$ *where:*

- *Ag is a finite, non-empty set of* agents*; and*
- $\leadsto \subseteq Ag \times Ag$ *is a binary* distrust *relation on Ag.*

When $i \leadsto j$ *we say that agent i distrusts agent j. We assume* $\leadsto$ *to be irreflexive, i.e., no agent i distrusts itself. Whenever i does not distrust j, we write* $i \not\leadsto j$*. So, we assume* $\forall i \in Ag$, $i \not\leadsto i$*. Call an agent i* fully trustworthy *if for all* $j \in Ag$*, we have* $j \not\leadsto i$*. Also, i is* trustworthy *if for some* $j \neq i$, $j \not\leadsto i$ *holds. Conversely, call i* fully trusting *if for no j,* $i \leadsto j$*. And i is* trusting *if for some* $j \neq i$, $i \not\leadsto j$*.*

Definition 2 *If* $S_1 = \langle Ag_1, \leadsto_1 \rangle$ *and* $S_2 = \langle Ag_2, \leadsto_2 \rangle$ *are two* ATF*s, we say that* S_2 *extends* S_1*, written* $S_1 \sqsubseteq S_2$*, if both* $Ag_1 \subseteq Ag_2$ *and* $\leadsto_1 \subseteq \leadsto_2$*.*

Example 1 *Consider the* ATF*s in Figure 1 and Figure 2. Vertices represent agents and if* $a \leadsto b$*, we represent this by an arrow from a to b. In the* ATF*s* S_1 *and* S_2*, distrust takes the pattern of a cycle. In particular, all agents are trustworthy (but no agent is fully trustworthy) and all agents are trusting (but none is fully trusting). In* S_3*, agent a is not trusting.* ATF S_4 *represents a situation where agents may be physically located linearly, and no agent trusts any of its neighbours. In* S_5*, agents a and d are fully trustworthy and c and d are both fully trusting. The society* S_8 *is like* S_2*, but now there is an additional agent d who is the only fully trustworthy agent. It distrusts b.*

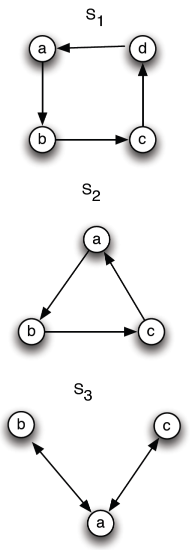


Fig. 1. Three simple ATFs

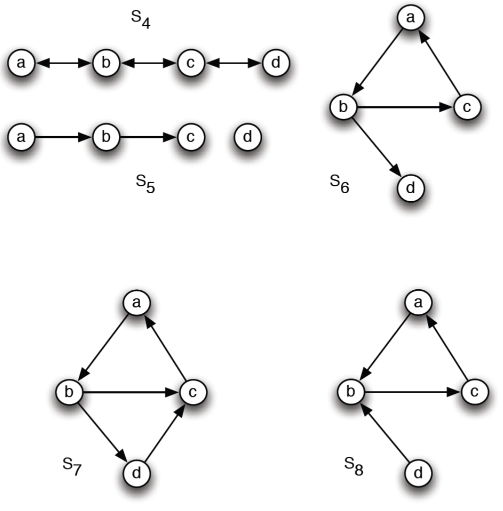


Fig. 2. Five ATFs for four agents

Coalitions with Trust A "coalition" is nothing other than a subset C of Ag. When forming a coalition, there are several ways to measure how much distrust there is among its members, or how trustable the coalition is with respect to the overall set of agent Ag.

Definition 3 *Given an* ATF $S = \langle Ag, \rightsquigarrow \rangle$, *a coalition* $C \subseteq Ag$ *is* distrust-free *if no member of C distrusts any other member of C. Note that the empty coalition and singleton coalitions* $\{i\}$ *are distrust-free: we call them trivial coalitions.*

Distrust freeness can be thought of as the most basic requirement for a *trusted* coalition of agents. It means that a set of agents has no internal distrust relationships between them. Since we assume $\rightsquigarrow$ to be irreflexive, we know that for any $i \in Ag$, the coalition $\{i\}$ is distrust-free, as is the empty coalition. A distrust-free coalition for S_5 in Figure 2 is, for example, $\{a, c, d\}$, and, while $\{b, c\}$ is distrust-free in S_3, society S_2 has no distrust-free coalitions other than the trivial ones.

Consider ATF S_5 from Figure 2. The coalition $C_1 = \{c, d\}$ is distrust-free, but still, they are not angelic: one of their members is being distrusted by some agent in Ag, and they do not have any justification to ignore that. Compare this to the coalition $C_2 = \{a, c, d\}$: any accusations about the trustworthiness of c by b can be neutralised by the fact that a does not trust b in the first place. So, as a collective, they have a defense against possible distrust against them.

Definition 4 *Let* ATF $S = \langle Ag, \rightsquigarrow \rangle$ *be given.*

- *An agent* $i \in Ag$ *is called* trustable *with respect to a coalition* $C \subseteq Ag$ *iff* $\forall y \in Ag((y \rightsquigarrow i) \Rightarrow \exists x \in C(x \rightsquigarrow y))$.

- *A coalition $C \subseteq Ag$ is a* trusted extension *of S iff C is distrust-free and every agent $i \in C$ is trustable with respect to C.*
- *A coalition $C \subseteq Ag$ is a* maximal trusted extension *(*mte) of S if C is a trusted extension, and no superset of C is one.

It is easy to see that if $i \in Ag$ is trustable with respect to some coalition C, then i is also trustable with respect to any bigger coalition $C' \supseteq C$. We will see that (maximal) trusted extensions are not closed under supersets, though.

Consider the two ATFs S_1 and S_2 of Figure 1, where distrust takes the form of a cycle. In S_2, the only coalitions C with respect to which a is trustable are the coalitions that have b as a member: indeed, agent c distrusts a, but c in turn is distrusted by b. However, there is no trusted extension of S_2, which is easily seen as follows. Suppose a would be in a trusted extension S. Since c distrusts a, there needs to be an agent in S that distrusts c. The only agent that qualifies would be b, but b and a cannot be at the same time in a distrust-free coalition.

Contrast this to the society S_1, where the distrust relation also forms a cycle, but where the two extensions are $\{a,c\}$ and $\{b,d\}$. It is easy to verify that those are also maximal trusted extensions. So the two cycles S_1 and S_2 demonstrate that a ATF can have several, or no maximal trusted extensions, respectively.

Note that we have $S_2 \sqsubseteq S_6 \sqsubseteq S_7$. We saw that S_2 has no maximal extensions, and inspection shows that S_6 has neither. However, S_7 does have a *mte*: $\{a,d\}$. This shows that an extended society can gain extensions, as soon as d starts to distrust c in S_6, agents a and d together have a good story why they would work together, and not with anybody else.

The concept of a *trusted extension* represents a basic and important notion for agents who want to rationally decide who to form a coalition with, basing their decisions on trust. In particular: *a trusted extension is composed of agents that have a rational basis to trust each other*.

Example 2 *Consider the* ATF *S_4 in Figure 2. According to [9], the process of forming a coalition is initiated by a single agent, say a, who will ask other agents in the society whom it believes trustworthy to join it in a coalition. Assume a is looking for two fellow agents to form a coalition with. According to the* ATF *S_4, a distrusts only b, therefore a can ask all the agents but b to form a coalition with it. Suppose a asks c and since c does not have a reason not to trust a, it accepts the invitation. We now have a temporary coalition C_1 formed by $\{a,c\}$. The process continues and a decides to ask d to join. Again, d does not distrust a so it accepts too. Therefore the final coalition C_1 is formed by $\{a,c,d\}$. All agents in C_1 are trusted by a and they all trust a.*

In this example, it is easy to see that c is not trusted by d and d itself is not trusted by c. Therefore if, for example, the agent starting the process of forming a coalition had been c, d would have refused to join the coalition with it. Similarly, if c had been asked by d to form a coalition with it, it would have refused. Nevertheless, following this kind of approach, which uses trust as a factor in forming coalition, the result would be a coalition where at least two out of its three members, c and d, do not trust each other.

Coalition stability is a crucial problem. If agents do not trust the other components of the coalition, they could break away from the alliance, playing a negative role on the stability. Therefore, trust plays an important role for coalition stability. *Trusted extensions* provide a simple method for the agents to find *coalitions* where all the members are satisfied with their components.

Weak and Strong Trust Maximal trusted extensions are a very interesting concept when considering forming trusting coalitions. As mentioned in Section 2, it is possible that a particular ATF has more than one *mte*. One could assume that all the agents in the maximal trusted extensions are equally trustworthy. For example, with regard to ATF S_1, in Figure 1 the maximal trusted extensions are $\{a,c\}$ and $\{b,d\}$. In this case, the agents appear respectively in only one maximal trusted extension. Now consider the ATF S_5, in Figure 2. Here the maximal trusted extensions are $\{a,c,d\}$ and $\{b,d\}$. Suppose we are trying to determine the status of two agents, a and d. One way to address this is to consider how many times a and d occur in the maximal trusted extensions. Agent d occurs in all the maximal trusted extensions, while a occurs in just one of them. Hence we can take this as evidence that d is somehow more "trustworthy" than a.

Definition 5 *Let* ATF $S = \langle Ag, \rightsquigarrow \rangle$ *be given. An agent* $i \in Ag$ *is* Strongly Trusted *if it is a member of* every mte. An agent $i \in Ag$ is *Weakly Trusted* if it is a member of at least *one mte.*

Therefore, returning to the ATF S_5 in Figure 2, agents a, b and c are weakly trusted and agent d is strongly trusted (and hence also weakly trusted). The notion of strongly and weakly trusted can help agents decide in those situation where there are large maximal trusted extensions but not all the agents are required for forming a stable coalition.

Personal Extensions In large societies, it is very unlikely that a single agent manages to interact with everyone in the society. For this reason, it has to rely on information given by others, about reputation of the agents it doesn't know. Reputation can be defined as the opinion or view of someone about something [11]. This view can be mainly derived from an aggregation of opinions of members of the community about one of them. However, it is possible that the agent doesn't trust a particular agent and it wants to discard its opinion. Therefore, when it comes to forming a coalition, the agents wants to consider only its personal opinion and the opinion of the agent it trusts, while still keeping the coalition distrust-free.

For example, suppose that an agent wants to start a project and it needs to form a coalition to achieve its goals. It wants to form a coalition composed only of agents it trusts and who have no distrust relations among them. To capture this intuition, we introduce two notions of *personal extensions*, which make it precise.

Definition 6 *Let the* ATF $S = \langle Ag, \rightsquigarrow \rangle$ *and* $a \in Ag$ *be given. Then* $C \subseteq Ag$ *is a Maximal Personal Extension generated by a, (notation:* $MPE(S,a)$*), if it is a* maximal trusted extension *of the* ATF $S' = \langle Ag', \rightsquigarrow' \rangle$, *where:*

- $Ag' = \{x \in Ag \mid x \not\rightsquigarrow a\}$*; and*
- $\rightsquigarrow' = (Ag' \times Ag') \cap \rightsquigarrow$.

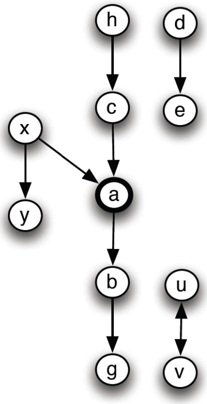
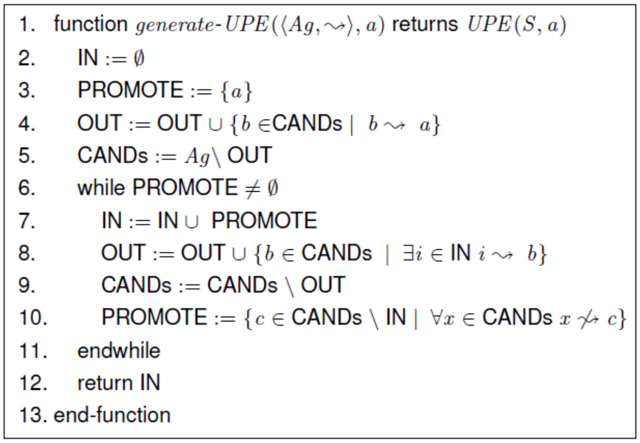


Fig. 3. S_{11}, an ATF for Personal Extension



```
function $generate\text{-}UPE(\langle Ag, \rightsquigarrow \rangle, a)$ returns $UPE(S, a)$
    IN := $\emptyset$
    PROMOTE := $\{a\}$
    OUT := OUT $\cup$ $\{b \in$CANDs | $b \rightsquigarrow a\}$
    CANDs := $Ag \setminus$ OUT
    while PROMOTE $\neq \emptyset$
        IN := IN $\cup$ PROMOTE
        OUT := OUT $\cup$ $\{b \in$ CANDs | $\exists i \in$ IN $i \rightsquigarrow b\}$
        CANDs := CANDs $\setminus$ OUT
        PROMOTE := $\{c \in$ CANDs $\setminus$ IN | $\forall x \in$ CANDs $x \not\rightsquigarrow c\}$
    endwhile
    return IN
end-function
```

Fig. 4. An algorithm for generating $UPE(S, a)$.

So, to form an *MPE*, we remove all agents that distrust a from Ag, and restrict the distrust relation to that set. Consider the ATF $S = \langle Ag, \rightsquigarrow \rangle$ in Figure 3. Suppose agent a wants to compute its personal extensions. Then, according to our definition, agents x and c will be discarded because they distrust a. Therefore, the *maximal trusted extensions* computed for this restricted set Ag' are $\{a, d, g, h, y, u\}$ and $\{a, d, g, h, y, v\}$. In general, we can say that, an agent b enters a maximal personal extension as long as everybody who distrusts it, is distrusted by someone who is accepted.

Although the extensions obtained this way are maximal (wrt set inclusion), this definition allows for more than one maximal personal extensions. Therefore, with regards to the example in Figure 3, agent a will have to choose between *two* personal extensions. However, without other information or additional criteria available to him, it wouldn't be able to make a justified decision. Therefore we introduce our second notion of personal extension, the *unique personal extension*.

Given an ATF $S = \langle Ag, \rightsquigarrow \rangle$, and an agent $a \in Ag$, the unique personal extension $UPE(S, a)$, we require, has the following properties:

1. $a \in UPE(S, a)$
2. $UPE(S, a)$ is unique
3. $UPE(S, a)$ is distrust free
4. there is a minimal set $OUT \subseteq Ag$, with the following properties, for all $x, y \in Ag$:
   (a) $x \rightsquigarrow a \Rightarrow x \in OUT$
   (b) $(y \in UPE(S, a) \ \& \ y \rightsquigarrow x) \Rightarrow x \in OUT$
   (c) $y \in UPE(S, a) \Leftrightarrow \forall z(z \rightsquigarrow y \Rightarrow z \in OUT)$

Loosely put: we add a to $UPE(S, a)$, and then we ensure that whoever distrusts or is distrusted by sombody in $UPE(S, a)$ is out, while $UPE(S, a)$ only accepts those agents as members that are at most distrusted by members of OUT.

We define $UPE(S, a)$ through an algorithm that generates it, from which the first three properties can be directly derived (see Figure 4). The algorithm works as follows.

Given an ATF $S = \langle Ag, \rightsquigarrow \rangle$, we take an agent $a \in Ag$, for whom we want to compute the unique personal extension $UPE(S,a)$, with the idea that this extension is conceived as iteratively computing sets of agents *IN* (the agents accepted in the process) using sets *OUT* (the agents that are rejected), *CANDs* (agents not in *OUT*) and, finally, a set *PROMOTE*: those agents from *CANDs* that stand the test that they are not distrusted by any agent in *CANDs* and go to *IN*. The properties $IN \subseteq CANDs$ and $CANDs = Ag \setminus OUT$ are *invariants* of the algorithm: they are both true at line 5, before entering the while-loop, and at line 10, at the end of the loop. Initially, nobody is *IN* (line 2 of Figure 4), but we put a, the agent whose personal extension we are computing, in *PROMOTE* (line 3), while the agents who distrust a are definitely *OUT*. Then, as long as there are agents to be promoted, do the following: mark the agents to be promoted as *IN* , (line 7), and make those agents that are distrusted by any agent that is *IN* definitely *OUT* (line 8). Then remove the agents that are *OUT* from the set of *CANDs* (line 9) and *PROMOTE* those agents that are candidates but not yet in if they are not distrusted by any candidate in *CANDs* (line 10). The agents remaining in *IN* after this process form the agent a's unique personal extension: $UPE(S,a)$.

Note that agents can be out for two reasons: first of all, they may distrust agent a (line 4), or they may themselves be distrusted by an agent that is in (line 8).

Theorem 1 *Let the* ATF $S = \langle Ag, \rightsquigarrow \rangle$ *and* $a \in Ag$ *be given. If we define* $UPE(S,a)$ *as the set IN returned by the function generate-*$UPE(S,a)$*, then* $UPE(S,a)$ *satisfies the four requirements set out above.*

Consider the example shown in Figure 3, for which we now calculate $UPE(S,a)$.

- Initially, *IN* is the emptyset, *PROMOTE* becomes $\{a\}$ and *OUT* becomes $\{x\}$, since x distrusts a;
- We then enter the while loop, during which in the first cycle, agent a enters *IN*, and *OUT* becomes $\{x,b\}$, since b is now distrusted by someone in *IN*. Everybody outside *OUT* is in *CANDs* and now the agents to be promoted are $\{d,g,h,y\}$: d and h are promoted since they are not distrusted by anybody, and g and y are promoted because the only agents that distrusted them (b and x, respectively), are now out.
- In the next cycle of the while-loop, the variable *IN* becomes $\{a,d,g,h,y\}$ and *OUT* is $\{x,b,c,e\}$ (the new members c and e are distrusted by the new *IN*-members h and d, respectively. The set *CANDs* now becomes $\{a,d,g,h,y,u,v\}$, and *PROMOTE* is now empty and the program terminates with $IN = \{a,d,g,h,y\}$.

Note that the two agents u and v are candidates throughout the algorithm and never become a member of *IN* or *OUT*. In general, we have that $UPE(S,a)$ is included in every personal extension generated by a set a. In fact, the two maximal personal extensions generated by a in the ATF of Figure 3 are $UPE(S,a) \cup \{u\}$ and $UPE(S,a) \cup \{v\}$. In $UPE(S,a)$, an agent b only enters if all agents distrusting b are already eliminated, while in the maximal personal extensions generated by a, we may allow some other agents, as long as everybody who distrusts them is distrusted by someone who is accepted. This yields to bigger coalitions, but, as shown, the result is not unique.

All personal extensions, maximal and unique, are *distrust-free*. However, since the distrust relationship is not symmetric, therefore if i distrusts j, it is not necessarily the case that j distrusts i, it can happen that one agent's personal extension is not *trusted* according with our definition 4. The agent preventing the extension from being trusted will be the agent who the personal extension belongs to. With respect to example in Figure 3, agent a is not *trustable*, as the agent distrusting him, agent x is not distrusted himself by any of the agent in the personal extension. However, for the purpose of the personal extension, this is not a problem because it represents the coalition that agent a would choose for himself. Clearly agent a considers himself trustworthy and agents who distrust him are not part of this coalition.

Theorem 2 *Let* ATF $S = \langle Ag, \rightsquigarrow \rangle$*, be given. Then*
if $(\exists a \in Ag \mid \forall i \in Ag \; i \not\rightsquigarrow a)$ *then* $(\forall j \in UPE(S,a), UPE(S,j) = UPE(S,a))$

Informally, Theorem 2 says that:

> *if the unique personal extension is computed by an agent who is distrust-free in the whole society, therefore it is fully trustworthy, then all the agents in that extension will have the same unique personal extension.*

Consider, for example, agent h's unique personal extension, from the ATF $S = \langle Ag, \rightsquigarrow \rangle$ in Figure 3. Following the algorithm in Figure 4, agent h's unique personal extension will be formed by $\{d, h, x, b\}$. We can notice that the unique personal extensions of agents d, x, and b are also $\{d, h, x, b\}$. If the agent is not distrusted in the society, then the other agents in its unique personal extension will share its vision of personal trustworthy coalition.

3 Aggregate Trust Measures

Abstract trust frameworks provide a social model of (dis)trust; they capture, at a relatively high level of abstraction, who (dis)trusts who in a society, and notions such as trusted extensions and personal extensions use these models to attempt to understand which coalitions are free of negative social views. An obvious question, however, is how the information presented in abstract trust frameworks can be *aggregated* to provide a single measure of how trustworthy (or otherwise) an individual within the society is. We now explore this issue. We present three aggregate measures of trust, which are given relative to an abstract trust framework $S = \langle Ag, \rightsquigarrow \rangle$ and an agent $i \in Ag$. Both of these trust values attempt to provide a principled way of measuring the overall trustworthiness of agent i, taking into account the information presented in S:

- *Distrust Degree*: This value ignores the structure of an ATF, and simply looks at how many or how few agents in the society (dis)trust an agent.
- *Expected trustworthiness*: This value is the ratio of the number of maximal trusted extensions of which i is a member to the overall number of maximal trusted extensions in the system S.
- *Coalition expected trustworthiness*: This value attempts to measure the probability that an agent $i \in Ag$ would be trusted by an arbitrary coalition, picked from the overall set of possible coalitions in the system.

These latter two values are related to solution concepts such as the Banzhaf index, developed in the theory of cooperative games and voting power, and indeed they are inspired by these measures [5].

Distrust Degree On the web, several successful approaches to credibility such as PageRank [2] use methods derived from graph theory to model credibility, which utilize the connections of the resource for evaluation. Several graph theoretic models of credibility and text retrieval [11] rely on the consideration of the in-degree of the vertex, that is the sum of the incoming edges of that particular vertex in a directed graph. The degree of the incoming edges is used to extract importance and trustworthiness. In our model, incoming edges are distrust relationships, therefore they represent a negative evaluation of a particular agent from the others in the society. Thus, measuring the in-degree of an agent in the society can give an indication how reliable (or unreliable) that agent is considered overall.

Formally, we call this value the *distrust-degree* for an abstract trust framework $S = \langle Ag, \leadsto \rangle$ and an agent $i \in Ag$, denoted as $\delta_i(S)$, and it is defined:

$$\delta_i(S) = \frac{|\{x \mid x \in Ag \ and \ x \leadsto i\}|}{|Ag|}.$$

This number provides us a measure of the reliability of the agent in the whole society. The higher the number of agents that distrust it, the less reliable that agent is considered to be.

However, as we mentioned before, a maximal trusted extension or, in general, a coalition C, according to our approach, is a set of agents who trust each other. Therefore, these agents may not be interested in the evaluation of the agents outside the coalition. They are more interested in a distrust degree relative to C. Hence, we define the following measure. The *coalition distrust-degree* for an abstract trust framework $S = \langle Ag, \leadsto \rangle$, a coalition C and an agent $i \in Ag$, denoted as $\delta_i^C(S)$, defined as:

$$\delta_i^C(S) = \frac{|\{x \mid x \in C \ and \ x \leadsto i\}|}{|C|}.$$

The coalition distrust degree provides a measure for the agents in C to select agents outside the trusted coalition, who they believe to be more reliable among the agents in the society. Agents in C can rank the agents outside using the value of the coalition distrust degree. In this way, it is possible to obtain an ordered list of the agents who the coalition consider less unreliable. The smaller the value of the coalition distrust-degree, the more reliable the agent is considered.

Expected Trustworthiness As we noted above, the expected trustworthiness of an agent i in system S is the ratio of the number of maximal trusted extensions in S of which i is a member to the overall number of maximal trusted extensions in the system S. To put it another way, this value is the probability that agent i would appear in a maximal trusted extension, if we picked such an extension uniformly at random from the set of all maximal trusted extensions. Formally, letting $mte(S)$ denote the set of maximal trusted extensions in $S = \langle Ag, \rightsquigarrow \rangle$, the expected trustworthiness of agent $i \in Ag$ is denoted $\mu_i(S)$, defined as:

$$\mu_i(S) = \frac{|\{C \in mte(S) \mid i \in C\}|}{|mte(S)|}.$$

Clearly, if $\mu_i(S) = 1$ then i is strongly trusted, according to the terminology introduced above, and moreover a is weakly trusted iff $\mu_i(S) > 0$.

From existing results in the argumentation literature on computing extensions of abstract argument systems [4], we can also obtain the following:

Proposition 1 *Given an* ATF $S = \langle Ag, \rightsquigarrow \rangle$ *and an agent* $i \in Ag$*:*

1. *It is #P-hard to compute* $\mu_i(S)$*.*
2. *It is NP-hard to check whether* $\mu_i(S) > 0$*.*
3. *It is co-NP-hard to determine whether* $\mu_i(S) = 0$*.*

For example, with respect to S_4 in Figure 2, the maximal trusted extensions are $\{a, c\}$ and $\{b, d\}$, therefore the expected trustworthiness of all the agents in the maximal trusted extensions is 0.5, since each of them is present in only one of the two maximal trusted extensions. However, if this society would have an additional agent e distrusted by nobody, the maximal trusted extensions would be $\{a, c, e\}$ and $\{b, d, e\}$, yielding a trustworthiness of agent e of 1, (it is a strongly trusted agent), while the expected trustworthiness of the other agents would be 0.5.

As an aside, note that the expected trustworthiness value is inspired by the *Banzhaf index* from cooperative game theory and voting theory [5].

Coalition Expected Trustworthiness There is one obvious problem with the overall expected trustworthiness value, as we have introduced above. Suppose we have a society that is entirely trusting (i.e., the entire society is distrust free) apart from a single "rogue" agent, who distrusts everybody apart from himself, even though everybody trusts him. Then, according to our current definitions, there is no maximal trusted extension apart from the rogue agent. This is perhaps counter intuitive. To understand what the problem is, observe that when deriving the value $\mu_i(S)$, we are taking into account the views of *all* the agents in the society – which includes every rogue agent. It is this difficulty that we attempt to overcome in the following measure. To define this value, we need a little more notation. Where $R \subseteq X \times X$ is a binary relation on some set X and $C \subseteq X$, then we denote by $restr(R, C)$ the relation obtained from R by restricting it to C:

$$restr(R, C) = \{(s, s') \in R \mid \{s, s'\} \subseteq C\}.$$

Then, where $S = \langle Ag, \rightsquigarrow \rangle$ is an abstract trust framework, and $C \subseteq Ag$, we denote by $S \downarrow C$ the abstract trust framework obtained by restricting the distrust relation $\rightsquigarrow$ to C:

$$S \downarrow C = \langle C, restr(\rightsquigarrow, C) \rangle.$$

Given this, we can define the *coalition expected trustworthiness*, $\varepsilon_i(S)$, of an agent i in given an abstract trust framework $S = \langle Ag, \rightsquigarrow \rangle$ to be:

$$\varepsilon_i(S) = \frac{1}{2^{|Ag|-1}} \sum_{C \subseteq Ag \setminus \{i\}} \mu_i(S \downarrow C \cup \{i\}).$$

Thus, $\varepsilon_i(S)$ measures the expected value of μ_i for a coalition $C \cup \{i\}$ where $C \subseteq Ag \setminus \{i\}$ is picked uniformly at random from the set of all such possible coalitions. There are $2^{|Ag|-1}$ coalitions not containing i, hence the first term in the definition.

$C \subseteq Ag \setminus \{i\}$	MPE in $C \cup \{i\}$	$\mu_i(C \cup \{i\})$
$\emptyset$	$\{a\}$	1
$\{b\}$	$\{a\},\{b\}$	0.5
$\{c\}$	$\{a,c\}$	1
$\{d\}$	$\{a,d\}$	1
$\{b,c\}$	$\{a,c\},\{b\}$	0.5
$\{b,d\}$	$\{a,d\},\{b\}$	0.5
$\{c,d\}$	$\{a,c\},\{a,d\}$	1
$\{b,c,d\}$	$\{a,c\},\{b,d\}$	0.5

$C \subseteq Ag \setminus \{i\}$	MPE in $C \cup \{i\}$	$\mu_i(C \cup \{i\})$
$\emptyset$	$\{b\}$	1
$\{a\}$	$\{a\},\{b\}$	0.5
$\{c\}$	$\{b\},\{c\}$	0.5
$\{d\}$	$\{b,d\}$	1
$\{a,c\}$	$\{a,c\},\{b\}$	0.5
$\{a,d\}$	$\{a,d\},\{b\}$	0.5
$\{c,d\}$	$\{b,d\},\{c\}$	0.5
$\{b,c,d\}$	$\{a,c\},\{b,d\}$	0.5

Table 1. Table showing the values of μ_i for $i = a$ (left) and $i = b$ (right) to calculate the Coalition Expected Trustworthiness with regard to example S_4 in Figure 2

Consider S_4 in Figure 2. In Table 1 we have shown a break down of all elements necessary to compute the Coalition Expected Trustworthiness for all the agents in S_4. Note that the value of $\mu_a(S_4)$ and $\mu_d(S_4)$ will be the same and the value of $\mu_b(S_4)$ and $\mu_c(S_4)$ will be the same as well. This is due to the particular shape of S_4. Therefore it is easy to notice that $\mu_a(S_4) = \mu_d(S_4) = 0.75$, while $\mu_b(S_4) = \mu_c(S_4) = 0.625$.

Note that in generalt the expected trustworthiness and the coalition expected trustworthiness differ. The coalition expected trustworthiness value arguably gives us a clearer overall idea of what the trustworthiness of an agent would be with respect to the maximal trusted extensions that can potentially be formed, therefore it offers a better insight in the trust issue related to the problem of forming coalitions.

4 Conclusions and Future Work

In this paper we have taken the notion of *distrust* as the key social concept. Based on this, we formulated several notions of mutually trusting coalitions. We have also presented techniques for how the information presented in our distrust model can be aggregated to produce individual measures to evaluate the trustworthiness of the agent with respect to the whole society or to a particular coalition.

We have at various times been talking about the notion of *stability* in *coalition formation*. There is, of course, another literature on stability in coalition formation: that of cooperative game theory – see [8] for details. The best-known notion of stability in cooperative game theory is the *core*: roughly, this solution concept considers a coalition to be stable if no subset of the coalition has any rational incentive to defect from the coalition, in the sense that they could earn more for themselves by defecting. Our model is not utility based, but there is nevertheless an interesting relationship between our notion of stability and that of cooperative game theory. In our view, a coalition is stable if no agent has any rational incentive to distrust any of the members. Future work might consider examining the role of our distrust models in coalition formation in more detail, perhaps in the context of coalition formation algorithms such as those recently proposed within the MAS community (see, e.g., [12]).

There are many potential directions for future work. We assume that the agents are willing to share their information about the trust they have in other agents. It would be interesting to devise some form of incentives for the agents to do so. Also, distrusted agent might not necessarily be trusted and we are investigating the possibility of combining both concepts in a more comprehensive framework. Moreover, our work is based on a boolean notion of trust. It would be intersting to add degrees of trust or distrust to allow for more detailed information. And finally, it would be interesting to introduce a measure of the abilities of the agents. This will allow for different perspectives to be analysed. By introducing this measure, we aim to make the framework more general, in order for it to be able to consider features other than only trust.

References

1. Breban, S., Vassileva, J.: Long-term coalitions for the electronic marketplace. In: Proceedings of the E-Commerce Applications Workshop, Canadian AI Conference (2001)
2. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Seventh International World-Wide Web Conference (WWW 1998) (1998), http://ilpubs.stanford.edu:8090/361/
3. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. AI 77, 321–357 (1995)
4. Dunne, P.E., Wooldridge, M.: Complexity of abstract argumentation. In: Rahwan, I., Simari, G. (eds.) Argumentation in Artificial Intelligence. SV (2009)
5. Felsenthal, D.S., Machover, M.: The Measurement of Voting Power. Edward Elgar: Cheltenham, UK (1998)
6. Griffiths, N., Luck, M.: Coalition formation through motivation and trust. In: Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems (2003)
7. Lei, G., Xiaolin, W., Guangzhou, Z.: Trust-based optimal workplace coalition generation. In: Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on. pp. 1 – 4 (2009)
8. Peleg, B., Sudholter, P.: Introduction to the Theory of Cooperative Games (second edition). SV (2002)
9. Qing-hua, Z., Chong-jun, W., Jun-yuan, X.: Core: A trust model for agent coalition formation. In: Natural Computation, 2009. ICNC '09. Fifth International Conference on. vol. 5, pp. 541 –545 (2009)
10. Rahwan, I., Simari, G.R. (eds.): Argumentation in Artificial Intelligence. SV (2009)
11. Sabater, J., Sierra, C.: Reputation and social network analysis in multi-agent systems. In: AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems. pp. 475–482. ACM, New York, NY, USA (2002)
12. Sandholm, T., Larson, K., Andersson, M., Shehory, O., Tohmé, F.: Coalition structure generation with worst case guarantees. AI 111(1–2), 209–238 (1999)
13. Wetzer, I.M., Zeelenberg, M., Pieters, R.: Never eat in that restaurant, i did!: Exploring why people engage in negative word-of-mouth communication. Psychology and Marketing 24(8), 661–680 (2007)

Design of a Reputation Mechanism for Virtual Reality: A case for E-Commerce

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

School of Computer Engineering, Nanyang Technological University, Singapore
{hfang1}@e.ntu.edu.sg
{zhangj,nadiathalmann}@ntu.edu.sg
[†]Department of Computing Science, University of Aberdeen, United Kingdom
{m.sensoy}@abdn.ac.uk

Abstract. The interest in 3D technology is growing both from academia and industry, promoting the quick development of 3D e-commerce (i.e. e-commerce systems in 3D virtual environments). In view of this, we propose a reputation mechanism particularly for 3D e-commerce. Supported by 3D technology and virtual reality, it consists of four major components: feedback provision based on human users' five senses, reputation computation based on feedback, 3D representation of computed reputation, and automatic decision making based on reputation. A user study is conducted to evaluate the necessity and value of our proposed reputation mechanism, confirming that users prefer 3D e-commerce with our proposed reputation mechanism over that with traditional reputation mechanisms. And, our proposed reputation mechanism can effectively ensure user's trust in the e-commerce system and simultaneously greatly promote user's trust in other users.

Categories: Models and mechanisms of reputation, Application studies

Keywords: Reputation Mechanism, Virtual Reality, Five Senses, Stereotype Trust, Feedback Alignment, User Study

1 Introduction

The Internet has become an inseparable part of our daily life nowadays. According to the Internet World Stats[1], the number of Internet users worldwide has reached 1.97 billion by the end of September 2010, accounting for almost 30 percent of the global population. Consequently, people are becoming more willing to shop online other than going to traditional solid shops. Unfortunately, current e-commerce systems only provide users with a simple, browser-based interface to acquire details of products and services. This kind of interfaces has been confirmed to be difficult for customers to use, and thus resulted in the low online shopping revenue [1]. One reason is the lack of effective interaction approaches, including communication channels and coordination methods between e-commerce systems and customers. Another more important reason is the limited understanding of social contexts, including social and behavioral issues,

[1] http://www.internetworldstats.com/stats.htm

among which trust is one of the most important issues. Besides, the design of current e-commerce systems is quite constrained and not appealing.

On another hand, 3D technology is gaining popularity. Forrest report [2] acclaims that "within five years, the 3D Internet will be as important for work as the web is today." A technology guru at Intel Corp also predicts that "the Internet will look significantly different in 5 to 10 years, when much of it will be three dimensional or 3D" [3]. Meanwhile, applications of virtual reality, such as immersing in 3D virtual communities, watching 3D movies and playing 3D games, are becoming part of ordinary life for people. 3D e-commerce, which is e-commerce systems in 3D virtual environments, has also attained growing interests both from academia and industry. It is one of the approaches proven to be effective in handling the problems in traditional e-commerce. As shown in Figure 1[1], the research gap between e-commerce and 3D technology or virtual reality is becoming smaller year by year. This partly explains the increasing research trend of 3D mall. Some industrial representatives of 3D e-commerce are IBM's VR-commerce program [4], Google lively project (http://www.lively.com), Second Life (http://secondlife.com), Active World (http://www.activeworlds.com), Twinity (http://www.twinity.com) and Virtual Shopping (http://virtualeshopping.com), etc.

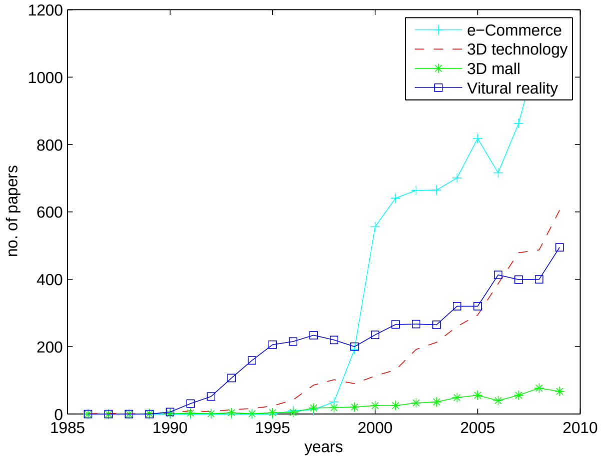


Fig. 1. Research Trend of E-Commerce, 3D Technology, Virtual Reality and 3D Mall

However, the same as traditional e-commerce systems, there are also inherited trust problems for 3D e-commerce. For one thing, some users may be dishonest. For example, sellers may not deliver the products as what they promised. For another thing, users may have different competency. For example, some sellers may produce only low quality products. Although there is the growing research interest towards 3D e-commerce, much research has been focused on either virtual reality technology adoption for e-commerce or behavioral science studies to confirm that 3D e-commerce environments like Active World can promote consumers' trust towards online shopping, without serious and quantitative consideration on how to construct effective trust and reputation

[1] Data was collected from the Web of Science on March 5, 2011

mechanisms in 3D e-commerce environments. For a few studies on designing reputation mechanisms for 3D e-commerce [5], they apply traditional reputation mechanisms where only simple numerical ratings, textual descriptions and 2D pictures are considered, overlooking the difference between 2D and 3D e-commerce environments.

To effectively address the trust issue in 3D e-commerce, we design a reputation mechanism specifically for 3D e-commerce environments. It is mainly built on buyers' feedbacks about their shopping experience with sellers and their subjective perceptions about products delivered by sellers. More specifically, in 3D environments, these kinds of feedback information can come from human users' five senses (vision, hearing, touch, smell and taste) enriched by virtual reality. We systematically study the four major steps of constructing the mechanism, namely feedback provision, reputation computation, reputation representation and decision making, by incorporating novel elements related to 3D e-commerce. We also conduct a detailed user study to compare our mechanism with traditional reputation mechanisms in 3D e-commerce environments. The results confirm that users prefer 3D e-commerce with our proposed reputation mechanism over that with traditional reputation mechanisms. Our mechanism can effectively ensure user's trust in the 3D e-commerce system and simultaneously greatly promote user's trust in other users. Our work thus represents a valuable first step of designing an effective reputation mechanism for promoting user participation in 3D e-commerce.

The rest of this paper is organized as follows. Section 2 provides an overview of related research on 3D e-commerce and reputation mechanisms. Section 3 illustrates our reputation mechanism for 3D e-commerce. The user study of comparing our mechanism with traditional reputation mechanisms in 3D environments is presented in Section 4. Finally, we conclude the current work and propose future work in Section 5.

2 Related Work

There are mainly two research directions on 3D e-commerce. The first direction concerns about adopting 3D technology and virtual reality into e-commerce, that is the construction of 3D e-commerce. This is also currently the major research towards 3D e-commerce. For example, Bogdanovych et al. [6] propose a mechanism called 3D E-Commerce Electronic Institutions and try to increase user's trust on e-commerce systems. The second direction mainly concerns about validating the effectiveness of 3D e-commerce in addressing the problems of traditional e-commerce. For example, Papadopoulou [7] demonstrates that a virtual reality shopping environment enables the formation of trust over conventional web stores, through a survey study based on a prototype virtual shopping mall. Nassiri [8] also explains the roles of 3D e-commerce environments in increasing user's trust and in improving profitability by the mechanisms such as Avatar appearance and Haptic tools. The research conducted by Teoh and Cyril [9] mainly focuses on the trust of 3D mall. They point out that presence and para-social presence assisted by virtual reality can affect trust, and users perceive the features of a 3D immersive online e-commerce store as being useful and practical but not a mere novelty. The weakness of the research mentioned above is that they focus only on enhancing trust through virtual reality. They do not consider how to improve trust in 3D e-commerce by designing effective trust and reputation mechanisms. This is the focus of our current work.

In recent years, a lot of research have been carried out on reputation mechanisms in traditional 2D e-commerce, and have achieved a huge success, while one of well known reputation systems is run by eBay (www.ebay.com). EBay's reputation system, also as one of the earliest online reputation systems, gathers feedbacks from buyers of each transaction in the simple form of numerical ratings together with a short text description. There are other successful commercial and live reputation systems [10], such as expert sites like Askme (www.askmecorp.com), products review sites like Epinions (www.epinions.com), and scientometrics related sites. However, there are only a few studies on designing reputation mechanisms specifically for 3D environments. Huang et al. [5] propose a reputation mechanism based on peer-rated reputations for 3D P2P game environments where the reputation of each user is computed based on other users' subjective opinions during their interactions, which is similar to eBay's reputation mechanism. It earned some advantages on reputation evaluation, storage, query and reliability, but no simulation has been conducted to validate its advantages. Its major weakness lies in the fact that there is no consideration of differences between 3D environments and 2D environments. In contrast, our reputation management makes good use of virtual reality to allow the provision of feedback information from human user's five senses. The other components of our reputation mechanism also follow such a design principle of fully utilizing the important features offered by virtual reality.

3 Reputation Mechanism for 3D E-Commerce

As mentioned in the previous section, current research focuses mainly on virtual reality technology adoption. Limited research on reputation mechanisms for 3D e-commerce however overlooks the differences between 2D environments and 3D environments. For a traditional reputation mechanism, buyer feedback often consists of only a positive, negative, or neutral rating, along with a short textual comment. Reputation of sellers is computed based on the ratings and perhaps those comments left by buyers, and is often in a form of a continuous numerical value. The computed reputation values will be used to make decisions for buyers on which sellers to do business with in the future.

Our reputation mechanism is specifically designed for 3D e-commerce environments. It is composed of four components: feedback provision, reputation computation, reputation representation and decision making. These components are supported by virtual reality and 3D technology, details of which will be explained in the subsequent subsections.

3.1 Feedback Provision

Feedback provision, as the key component of our reputation mechanism, tries to solve two major problems: what kind of user feedbacks to collect and how to collect feedbacks in 3D e-commerce environments. There are five senses - vision, hearing, touch, smell and taste, which express the *subjective perceptions* of human being. People have the ability to sense the environment and objects with these five senses, and further provide themselves better understanding of the environment. 3D e-commerce is a virtual environment generated by computer and other tools, such as head-mounted displays, headphones, and motion-sensing gloves, to enable users to feel realism through interaction that simulates five human senses. In traditional e-commerce mechanisms, only vision is regularly incorporated in simple forms like 2D pictures and textual descriptions. As human users' perception of an environment is influenced by all the sensory inputs, in order to accurately and completely express user's experience, all the five senses should be well expressed. With the development of virtual reality and augmented reality, the perception of human users not only can be realistically simulated, but also can be expanded by using instruments like 3D Glasses.

Five Senses: **Vision** is the ability to interpret information of what is seen from the environment, and can be expressed in the form of 3D pictures and videos in virtual reality. Therefore, in 3D e-commerce, buyers can present the real product they purchased in the form of 3D picture or animation with less distortion. Users can view the 3D object from various angles, which is more persuasive and vivid than simple 2D pictures or textual descriptions. **Hearing** is the ability to perceive sound from the environment, and can be simulated by auditory displays. Same as vision, there have been numerous works on auditory research. In 3D e-commerce, some characteristics such as tone quality of digital products are more appropriate to be presented in the form of audio. Audio is able to contain plentiful information at a time, and relatively favored and easily accepted by human users. In this sense, it is necessary to collect this kind of information. **Touch** is one of the sensations processed by the somatosensory system, and has been known in the physical world to increase initial trust. As a major part of research in virtual reality, it focuses on scanning the behaviors of objects in the physical world and incorporating similar behavior into virtual objects [11]. We have previously done some research on touching textile [12]. Touch perception can be simulated using instruments like Haptic device. Virtual touch can be supported in 3D e-commerce so that buyers can measure the characteristics of different materials and attach touch information to reputation feedback as guidance for other buyers. **Taste** refers to the ability to detect the flavor of substances such as food and minerals. Humans receive tastes through sensory organs called taste buds. The sensation of taste traditionally consists of some basic tastes such as sweetness, bitterness, sourness and saltiness. Taste can also be implemented in virtual environments. Iwata et al. [13] design a food simulator to simulate the multi-modal taste of food through a combination of chemical, auditory, olfactory and haptic sensation. Through this simulator, buyers can provide experience about the taste of products they purchase online. **Smell** refers to the ability to perceive odors. In 3D environments, devices like the olfactory display can be applied to generate various odors and deliver them to user's nose. For the purpose of presenting odors with a vivid sense of reality, the olfactory display, which has already been applied to 3D games and movies, is expected to generate realistic smells relevant to specific environments or scenes [14]. In 3D e-commerce, they can be realistic smells related to specific products such as fresh smell of fruits. Buyers can then sense a product's real smell through other buyers' feedbacks instead of textual descriptions about smells.

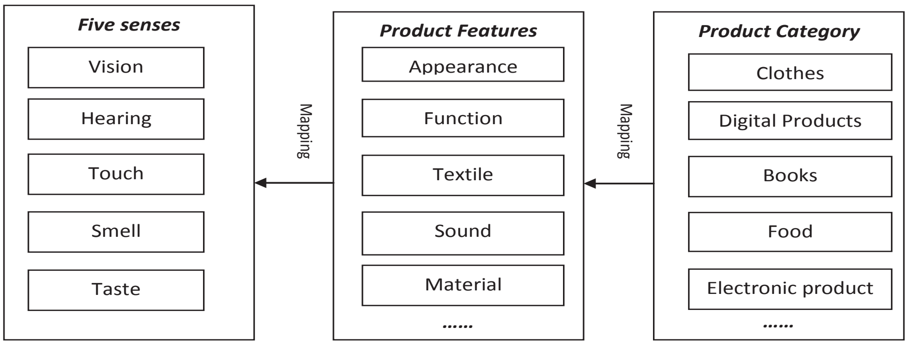


Fig. 2. Feedback Provision based on an Five-Sense Oriented Approach

Five-Sense Oriented Feedback Provision: As illustrated above, while concerning about buyers' historical experience with one seller, feedbacks can be expressed as human perceptions about the products and transaction experience. These subjective perceptions can be simulated by virtual reality. Therefore, towards 3D e-commerce environments, we propose a five-sense orientated approach to implement feedback provision as part of our reputation mechanism. The detail of the approach is illustrated in Figure 2. Consider a 3D e-commerce community providing products of different categories. According to the five-sense orientated approach, a product may belong to some specific product categories such as "Clothes" or "Books". Products in the same category have some common product features, such as "Appearance" and "Textile". Each product feature can be presented by some of the five senses - vision, hearing, touch, smell and taste simulated by virtual reality as mentioned earlier. Thus, given a product, the necessary senses will be simulated in feedback. For example, a user has purchased a sweater from a seller in a 3D e-commerce system. For feedback provision, the buyer can provide a 3D avatar model to express the appearance of the sweater sold by the seller. Besides, the touch feedback can also be simulated to show the textile and material used to make this sweater. Such information shared among buyers can be compared with the 3D avatar model of the product provided by the seller to compute reputation of the seller.

3.2 Trust/Reputation Computation

Here, we assume that each buyer (i.e., its agent) can produce a feedback for the product delivered by a seller. This feedback is based on the five senses of the consumers and represented using an *ontology* [15, 16] that contains a rich set of concepts, properties and individuals to represent the perceptions of the buyers. This representation allows decidable reasoning over the feedbacks using off-the-self ontology reasoners [17]. The seller has a description of the product in the virtual reality setting. Hence, a consumer receiving this description can use his five senses to evaluate the product via the simulated reality (e.g., using olfactory display). However, the actual product delivered by the seller may be different than the described product. In most of the existing models, to evaluate the trustworthiness of sellers, consumers use the information about their past transactions with these sellers. However, a consumer may not have enough number of historical information about many of the sellers. This disallows him to evaluate a seller based on his direct interactions. In such situations, we may evaluate the trustworthiness of sellers based on the available information such as personal features of sellers (e.g., location) or information from other consumers (i.e., their feedbacks).

Exploiting Features of Sellers To estimate trustworthiness of sellers, we can use a stereotype-based trust model [18] based on a rich set of seller features in virtual environments. That is, using the personal interactions with previously encountered sellers, the buyer can derive some rules that allow him to characterize other sellers with specific features as less or more trustworthy. Existing stereotypical trust models learn rules using the features of the sellers and their products. For instance, in [18], these rules are learned using regression trees. Each rule maps sellers with specific features onto a trust value in the range [0,1]. However, [18] assumes only numerical features, while many of the features in real-life settings are nominal (i.e., categorical). The reason behind this assumption is the fact that decision or regression trees cannot make generalization or induction over nominal values. To address this issue, we can extend C4.5 decision trees [19] by exploiting domain knowledge during tree induction. That is, nominal values of attributes are generalized using taxonomy of attribute values, which can easily be derived from domain ontologies.

Figure 3 shows an example decision tree built based on a buyer's interaction history with sellers and the feedback from other buyers, using C4.5 algorithm with domain knowledge. From this decision tree, the buyer creates stereotypes such as 'seller from Europe cannot provide products with raw taste, but they may provide products with barbecue smell'. Hence, for a seller who sends a description of a product with a raw taste, the buyer may not trust the seller. As the buyer experiences more about sellers and receives new feedbacks, he updates his stereotypes based on the new information by building new decision trees.

Note that, personal interactions with sellers are very costly; hence we assume that buyers have little direct information with the sellers. However, once the buyer interacts with a specific seller, he can build a fine grained trust evaluation of the seller based on these personal interactions. On the other hand, in the absence of personal interactions or feedbacks about a specific seller, stereotypes help the buyer to evaluate the seller. In other words, stereotypes are used to bootstrap trust towards a specific seller, but then the trustworthiness of the seller is computed using other methods based on feedback about and direct interactions with this specific seller. For instance, subjective logic [20] and the Bayesian network-based trust model [21] can be used to compute the trustworthiness based on the evidence about the seller.

Exploiting User Feedbacks Existing e-commerce systems like e-bay[1] allow consumers provide feedbacks in the form of ratings and reviews. The ratings are aggregated by the system to compute reputation of the sellers. Then, the reputation of sellers guides consumers while deciding on a specific seller among alternative. We believe that similar ideas can also be used in virtual reality environments. That is, to be sure that a specific

[1] http://www/ebay.com

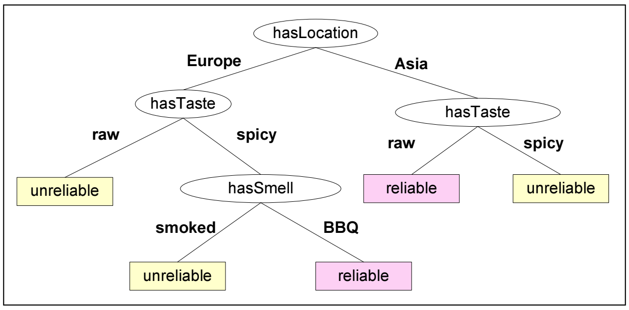


Fig. 3. A Decision Tree Example to Derive Simple Stereotypes

seller will provide the described product, the buyer may collect feedbacks of other buyers about the *same* seller. However, these feedbacks should be more expressive than ratings; they should contain context, sensory information provided by the seller before the transaction, and *subjective* evaluation of the actual product by the buyer. The provided sensory information, context and evaluation of the product can be mined using pattern recognition techniques [22] to learn critical information about the correlations of the advertised product features and the actual ones. The learned correlations can be exploited to reason about the reliability of sellers, given the sensory information they provide for a specific product. Here, an important problem is the subjectivity of the evaluations in users' feedbacks. Evaluations based on five senses such as tactile sensations are subjective. This means that a product evaluated as *too soft* by a user can be evaluated as *adequately soft* by another consumer. This brings the necessity of aligning subjective evaluations in feedbacks.

3.3 3D Visualization for Reputation Representation

Visualization is used to present reputation results of users. Traditional reputation mechanisms use visualization of 2D objects such as a simple rating score or characteristics descriptions in the form of text or 2D pictures, which is far from being effective and provides only limited information. We apply a 3D visualization approach, aiming at presenting a rich set of reputation related information in an appealing and natural way. In this way, users will be assisted to make more informed decisions and their trust in the reputation mechanism will be increased. 3D visualization to present reputation should follow some general principles and visualization requirement [23]. First, it should support users to achieve self-efficacy. Each user has an attractive reputation model, which can be built and enhanced further with the growing reputation. The growing process should be dynamic and be expressed in real time with the assistance of the time dimension. Secondly, the reputation of users should be easily recognized that there is a common criteria for reputation comparison. Thirdly, the visualization should support micro and macro reading. It refers to that user's overall reputation value can be easily identified. The details of user's reputation, such as reputation of specific product category or characteristics, should be displayed clearly.

3.4 Decision Making

Since a large number of sellers provide many similar products, it may take a lot of time for buyers to browse and search for the most suitable sellers. Our reputation mechanism will provide recommendations to buyers according to the computed reputation of sellers as well as buyers' preferences. For example, some risk-taking buyers may prefer low price of products and be willing to accept doing business with relatively low reputation of sellers. Some other buyers may care more about sellers' reputation.

4 User Study

In this section, we present a user study on comparing our proposed reputation mechanism with traditional reputation mechanisms in the same environment of 3D e-commerce. Since reputation computation and decision making are invisible to users, our study is concentrated on the feedback provision and reputation representation components.

4.1 Design of the Study

The comparison was based on two criterions. One is called "institutional trust" referring to user's trust in the mechanism, while the other is called "interpersonal trust" referring to user's trust in other users with the existence of reputation mechanisms. We measure the two kinds of trust by the framework of general trust - benevolence, competence, integrity and predictability [24]. Based on this guidance, a questionnaire survey is conducted. Figure 4 presents the overall structure of the questionnaire.

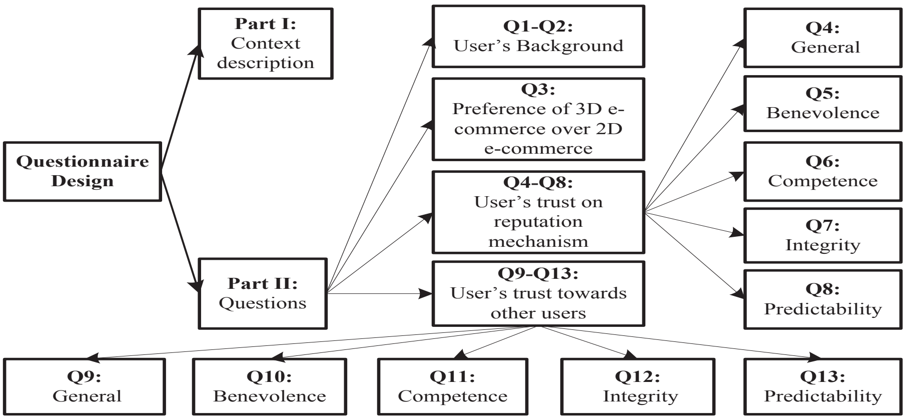


Fig. 4. Questionnaire Design for Data Collection

The questionnaire is divided into two main parts: context description part, which provides users the detailed description of our reputation mechanism and traditional reputation mechanism within 3D e-commerce environments; and questions part, consisting of 13 questions in total. In the context description, participants are presented with a set of images about what they will experience in the 3D e-commerce environment with our proposed reputation mechanism and that with the traditional reputation mechanisms. Besides, one researcher is responsible for the Q&A part in the process of questionnaire filling. Regarding the questions, Q1 and Q2 ask for the information of participant's background, including gender, age, nationality, current residency and online shopping background; Q3 aims to study user's preferences on 3D e-commerce versus 2D e-commerce; Q4-Q8 focus on studying user's trust on reputation mechanisms, referring to general trust, benevolence, competence, integrity and predictability of reputation mechanism respectively. Some examples are "Do you agree that compared with traditional reputation mechanisms, the proposed reputation mechanism provides you with more confidence in believing that 3D e-commerce is well-organized and the stores are benevolent to their customers?" and "Do you agree that the proposed reputation mechanism performs better in reducing fraud behaviors than traditional reputation mechanisms?"; Q9-Q13 try to explore user's trust in other users with the reputation mechanisms, and the structure is similar to Q4-Q8. The answers for each question can be chosen from the following five levels: "5-Totally agree", "4-Partially agree", "3-Neither Agree nor Disagree", "2-Partially disagree" and "1-Totally disagree".

A total of 40 subjects with the average age of 24 years old participated in the study. They were selected based on the stratified random sampling methods with respect to their gender and current residency. 21 of them are males. 21 of them are currently living in Asia, and 19 of them in America. Besides, all of them are experienced Internet users, but only 14 of them are within technology background, while 26 of them with the background of social science, management or related. 38 of them have purchased products online at least once a year, while 30 of them at least twice a year. The e-commerce systems they went shopping most often are Taobao (www.taobao.com), Amazon and eBay. One point should be emphasized here is that since the 3D E-Commerce per se is quite revolutionary, this study mainly focuses on the young generation mostly within the age of 22 years old to 26 years old, who are believed to be the major participants of 3D e-Commerce. The basic statistical information about the participants is summarized in Table 1 and 2. In addition, 26 (65%) of participants preferred 3D e-commerce over 2D e-commerce, while only 5 of them are willing to stay at 2D e-commerce sites, and 9 of them hold neutral attitude towards the preference of 3D e-commerce and 2D e-commerce.

Table 1. Statistical Information about the Participants I

	Gender		Nationality		Current Residency		Often Shopping Site		
	Male	Female	Asian	American	Asia	America	Taobao	Amazon+eBay	Others
Counts	19	21	24	16	21	19	16	17	7
Percents	47.5%	52.5%	60%	40%	52.5%	47.5%	40%	42.5%	17.5%

Table 2. Statistical Information about the Participants II

	Technology Background		Age Diversity					Attitude of 3D E-Commerce		
	Yes	No	18-21	22-23	24	25-26	27	Positive	Neutral	Negative
Counts	14	26	3	14	11	11	1	26	9	5
Percents	35%	65%	7.5%	35%	27.5%	26.5%	2.5%	65%	22.5%	12.5%

4.2 Data Analysis and Discussion

According to the trust framework of McKnight and Chervany [24], a good reputation mechanism promoting high trust of users should also assure users' beliefs such as benevolence, competence, integrity and predictability towards the reputation mechanism. Accordingly, a high degree of one perspective of the trust framework should also indicate a high degree of other perspectives. Based on these criterion and the collected data, we compute the pairwise correlation between trust and its four perspectives - benevolence, competence, integrity and predictability. Firstly, trust value of each participant is computed as the average value of Q4 and Q9. In the similar way, benevolence, competence, integrity and predictability values of each participant are computed according to participants' answers to Q5 and Q10, Q6 and Q11, Q7 and Q12, and Q8 and Q13 respectively. Each value is referred to participant's preference of our proposed reputation mechanism over traditional mechanisms. Then, the correlation analysis among each factor is conducted (See Table 3). By viewing the coefficient values, we find that trust is relatively highly correlated with each perspective (coefficients are all around 0.7000), especially for the correlation between trust and predictability (0.7449), indicating that people believe that 3D e-commerce with our proposed reputation mechanism would be competitive in the e-commerce market compared with that with the traditional reputation mechanisms. Additionally, the four perspectives are also relatively highly correlated with each other, which confirms that the trust framework in [24] can be applied to reputation mechanisms in 3D e-commerce.

In order to comprehensively compare our proposed reputation mechanism with traditional reputation mechanisms, we explore these 40 participants' evaluation towards the four perspectives of trust typology with respect to both their trust in the reputation mechanism (Institutional trust) and their trust in other users (Interpersonal trust). For Q4-Q13, the answers of "Totally Agree" or "Partially Agree" is treated as positive evaluation of our proposed reputation mechanism, "Neither Agree nor Disagree" as neutral evaluation, and "Partially Disagree" or "Totally Disagree" as negative evaluation. Table 4 presents the participants' specific evaluations (positive, neutral or negative) of each perspective concerned with each kind of trust regarding our reputation mechanism compared to those of conventional reputation mechanisms.

Table 3. Correlation between Trust related Variables

Variables	**Trust**	**Benevolence**	**Competence**	**Integrity**	**Predictability**
Trust	1.0000				
Benevolence	0.6970	1.0000			
Competence	0.6950	0.5939	1.0000		
Integrity	0.6985	0.7279	0.6241	1.0000	
Predictability	0.7449	0.7494	0.6441	0.6197	1.0000

User's Trust in the Mechanism According to the results in Table 4, to sum up, most (72.5%) of the participants showed stronger (institutional) trust in 3D e-commerce with our reputation mechanism than that with the traditional reputation mechanisms. In most of the participants' belief, our proposed reputation mechanism performs better in reducing fraud behavior (competence), provides them more confidence to believe in the 3D e-commerce (benevolence), and 3D e-commerce with our proposed reputation mechanism has greater possibility to achieve success (predictability) in the fierce competition.

Table 4. User Evaluation of our Reputation Mechanism over Traditional Reputation Mechanisms

Dimension		Positive		Neutral		Negative	
		Counts	Percents	Counts	Percents	Counts	Percents
User's trust in the mechanism	General	29	**72.5%**	3	7.5%	8	20%
	Benevolence	24	**60%**	8	20%	8	20%
	Competence	27	**67.5%**	10	25%	3	7.5%
	Integrity	17	**42.5%**	11	27.5%	12	30%
	Predictability	23	**57.5%**	8	20%	9	22.5%
User's trust in other users	General	23	**57.5%**	8	20%	9	22.5%
	Benevolence	20	**50%**	7	17.5%	13	32.5%
	Competence	25	**62.5%**	6	15%	9	22.5%
	Integrity	16	**40%**	12	30%	12	30%
	Predictability	27	**67.5%**	8	20%	5	12.5%

User's Trust in Other Users For the interpersonal trust, compared to traditional reputation mechanisms, users mostly hold a positive attitude towards our reputation mechanism. They are more confident that other users in our reputation mechanism are more trustworthiness (57.5%), while sellers would not only care more about buyers (50%) and more likely meet the quality requirement of the products as expected (62.5%), but also be more consistent with their behavior (67.5%) over time.

What should be noted is the integrity perspective both for institutional trust and interpersonal trust. Integrity refers to that sellers always provide high quality products and buyers always give truthful feedbacks. The integrity values of this study, although still positive, are relatively smaller (42.5% and 40%) compared to others, partly indicating that users worry about online shopping. Through interviewing the participants who expressed negative or neutral attitude towards our reputation mechanism, we found that they were just reluctant to use 3D e-commerce based on the technology limitations, but had less concern about reputation mechanisms.

Cultural Differences In addition, based on the user evaluation, the cultural differences between subjects living in Asia (mostly living in Singapore) and subjects living in America was also evaluated and the result was shown in Table 5. It demonstrates that, on the whole, both of them prefer our proposed reputation mechanism over traditional reputation mechanism, regarding the positive percents and negative percents. However, it should also be noticed that People living in Asia generally hold much more confident of our proposed reputation mechanism than people living in America. This can be explained that virtual reality has been greatly developed in Singapore and has many

Table 5. Comparison of People's Attitude towards our Reputation Mechanism over Traditional Reputation Mechanisms in Asia and America

Dimension		Positive		Neutral		Negative	
		Asia	America	Asia	America	Asia	America
User's trust in the mechanism	General	90.4%	52.6%	0%	15.8%	9.5%	31.6%
	Benevolence	76.2%	42.1%	14.3%	26.3%	9.5%	31.6%
	Competence	76.2%	57.9%	14.3%	36.8%	9.5%	5.3%
	Integrity	61.2%	**21.1**%	19%	36.8%	19%	**42.1**%
	Predictability	57.1%	57.9%	23.8%	15.8%	19%	26.3%
User's trust in other users	General	66.7%	47.4%	23.8%	15.8%	9.5%	36.8%
	Benevolence	57.1%	42.1%	19%	15.8%	23.8%	42.1%
	Competence	76.2%	47.4%	14.3%	15.8%	14.35%	31.6%
	Integrity	42.8%	36.8%	33.3%	26.3%	23.8%	36.8%
	Predictability	85.7%	47.4%	9.5%	31.6%	4.8%	21.1%

realistic applications, such as Virtual Singapore[1] and 3D Virtual World for 2010 Youth Olympic Games[2], while for America, it already has profound and mature development of traditional e-commerce websites, such as Ebay and Amazon, and the applications of 3D virtual world is relatively weak compared to those in European and some Asian countries. More cultures diversity, especially the attitude of people living in European, should be included in the further research.

5 Conclusion and Future Work

This paper proposes a reputation mechanism for 3D e-commerce by systematically studying the four steps of constructing reputation mechanisms, namely, feedback provision, reputation computation, reputation representation and decision making. We incorporate novel elements of 3D technology and virtual reality into these main steps. For feedback provision, a five-sense orientated approach is applied to provide buyers' feedbacks of products they have purchased in the form of five human senses simulated by virtual reality. For reputation computation, a multi-dimensional trust model and a stereotype-based approach may be applied to compute the reputation of sellers. 3D visualization is used to present computed reputation values. The proposed reputation mechanism can also effectively help users make purchase decisions. A user study is conducted to compare our mechanism with traditional reputation mechanisms in 3D e-commerce environments. The questionnaire survey with a stratified sampling method mainly focuses on user's trust in the mechanism (institutional trust) and user's trust in other users (interpersonal trust) respectively based on the four perspectives of trust typology - benevolence, competence, integrity and predictability. The findings illustrate that: (a) users prefer shopping in 3D e-commerce with our proposed reputation mechanism over that with traditional reputation mechanisms; (b) compared with traditional

[1] http://www.singaporevr.com/

[2] http://www.singapore2010odyssey.sg/

reputation mechanisms, our reputation mechanism can not only effectively ensure user's trust in the mechanism, but also greatly promote user's trust in other users.

Our current work represents an important initial step for confirming the necessity and value of our proposed reputation mechanism. For future work, we will first develop a concrete reputation computation method for our reputation mechanism and implement a 3D visualization scheme for reputation representation. A prototype of our reputation mechanism will be built to further study user's responses to 3D e-commerce with our proposed reputation mechanism, and more comprehensive user study, considering age diversity, shopping background and cultural differences, will be conducted.

6 Acknowledgement

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

References

1. Hoffman, D.L., Novak, T.P., Peralta, M.: Building consumer trust online. Communications of the ACM **42** (1999) 80–85
2. Drive, E.e.a.: Getting real work done in virtual worlds. Forrest Research (2008) `http://www.forrester.com/rb/Research/getting_real_work_done_in_virtual_worlds/q/id/43450/t/2`.
3. Gaudin, S.: Intel guru says 3-d internet will arrive within five years. Computer World (2010) `http://www.computerworld.com/s/article/9175048/Intel_guru_says_3_D_Internet_will_arrive_within_five_years`.
4. Mass, Y., Herzberg, A.: Vrcommerce- electronic commerce in virtual reality. In: Proceedings of the 1st ACM Conference on Electronic Commerce. (1999) 103–109
5. Huang, G.Y., Y., H.S., R., J.J.: Scalable reputation management for p2p mmogs. In: Proceedings of the International Workshop on Massively Multiusers Virtual Environment. (2008)
6. Bogdanovych, A., Berger, H., Simoff, S., Sierra, C.: Narrowing the gap between humans and agents in e-commerce: 3d electronic institutions. In: Proceedings of the 6th International Conference on Electronic Commerce and Web Technologies (EC-Web). (2005) 128–137
7. Papadopoulou, P.: Applying virtual reality for trust building e-commerce environment. Virtual Reality **11** (2007) 107–127
8. Nassiri, N.: Increasing trust through the use of 3d e-commerce environment. In: Proceedings of the ACM symposium on Applied computing. (2008) 1463–1466
9. Teoh, K. K.and Cyril, E.U.: The role of presence and para social presence on trust in online virtual electronic commerce. Journal of Applied Sciences **16** (2008) 2834–2842
10. Josang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decision Support Systems **43** (2007) 618–644
11. Pai, D.K.e.a.: Scanning physical interaction behavior of 3d objects. In: Proceedings of the 28th Annual Conference on computer graphics and interactive techniques. (2001)
12. Magnenat-Thalmann, N., Volino, P., Bonanni, U., Summers, I.R., asco, M.B., Salsedo, F., Wolter, F.E.: From physics-based simulation to the touching of textiles: the haptex project. The International Journal of Virtual Reality **6** (2007) 35–44
13. Iwata, H., Yano., T., Uemura, T., Moriya, T.: Food simulator: A haptic interface for biting. In: Proceedings of Virtual Reality (VR). (2004) 51–57

14. Brkic, B.R., Chalmers, A.: Virtual smell: Authentic smell diffusion in virtual environments. In: Proceedings of the 7th International Conference on Computer Graphics, Virtual Reality and Interaction in Africa. (2010) 45–52
15. W3C OWL Working Group: OWL 2 Web ontology language document overview (2009) http://www.w3.org/TR/owl2-overview.
16. Kern, S.: Olfactory Ontology and Scented Harmonies: on the History of Smell. The Journal of Popular Culture **7** (1974) 816–824
17. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical OWL-DL reasoner. Web Semant. **5** (2007) 51–53
18. Burnett, C., Norman, T.J., Sycara, K.: Bootstrapping trust evaluations through stereotypes. In: Proceedings of the Eighth International Conference on Autonomous Agents and Multiagent Systems. (2007)
19. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
20. Jøsang, A., Bhuiyan, T.: Optimal trust network analysis with subjective logic. In: Proceedings of the International Conference on Emerging Security Information, Systems and Technologies. (2008) 179–184
21. Wang, Y., Vassileva, J.: Bayesian network-based trust model. In: Proceedings of the International Workshop on Trust, Privacy, Deception and Fraud in Agent Systems. (2003) 372–380
22. Chen, C.: Handbook of Pattern Recognition and Computer Vision (4th Edition). World Scientific Publishing Co. Pte. Ltd (2010)
23. Erickson, T.: Designing visualizations of social activity: six claims. In: CHI'03 Extended Abstracts on Human Factors in Computing Systems. (2003) 846–847
24. McKnight, D., Chervany, N.: What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. International Journal of Electronic Commerce **6** (2001) 35–59

On the classification of emotions, and its relevance to the understanding of *trust*

Andrew J.I. Jones[1,2] and Jeremy Pitt[2]

[1]Department of Informatics, King's College London, UK
[2]Department of Electrical & Electronic Engineering,
Imperial College London, UK

Abstract. We outline some principal features of Ingmar Pörns modal-logical characterisation of some types of emotions, and indicate how that highly systematic approach might be applied to the analysis of trust and shame. We then suggest ways in which this approach might provide points of departure for further work: comparing the logical taxonomy with psychophysiological taxonomies; investigating responses, such as forgiveness, to the violation of trust; and how this logical taxonomy can be used as a platform for a cognitive agent architecture.

Keywords: Emotions, Trust, Shame, Forgiveness.

1 Introduction

We are interested in the question of whether some fundamental components might be identified, in terms of which a formal classifcation of types of emotions might be constructed. We explain, in Section 2 our reasons for thinking that a positive answer to that question can be given, and offer some conjectures regarding the analysis of *trust* and *shame*. Section 3 indicates some lines of further inquiry to which this formal taxonomy might be applied: comparing the logical taxonomy with psychophysiological taxonomies; investigating responses, such as forgiveness, to the violation of trust; and how this logical taxonomy can be used as a platform for a cognitive agent architecture.

2 A Formal Characterisation of Emotion

In a little known paper, published in 1986 [1], Ingmar Pörn outlined a formal-logical characterisation of some types of emotions. A key premise in his approach was that each emotion exhibits what he called *double intentionality*. For instance, if x hopes that the train arrived late, x "...wishes it to be the case that the train arrived late and he believes that it is possible that it did" (op. cit., p. 205). The emotion is of type *hope*, and the object of the emotion, in this example, is the state of affairs that the train arrived late; the double intentionality pertains to what x wishes and what x believes – respectively the *volitional* and *epistemic* aspects of the intentionality. Pörn sought to define a set of primary or atomic emotional types in terms of their respective volitional and epistemic components.

2.1 The Epistemic Component

In regard to the epistemic aspect Pörn made the further assumption that the key notion was that of certainty. He observed that such emotions as hope, fear and anxiety seem to be incompatible with certainty, whereas some other emotions – he here mentions regret, despair, anger and shame – seem to require it. Adopting the idea that 'x is certain that p' may be interpreted as 'x believes that he knows that p', he represented 'x is certain that p' as the modal-logical expression B_xK_xp simplifying this to BKp on the assumption that we consider one and the same individual in both subscript positions.

Following Pörn, we may now apply the kinds of basic moves employed in the theory of normative positions (see, e.g., [2]) in order to generate the set of atomic epistemic (certainty) positions. We here assume that the belief and knowledge modalities are both normal, and that their logic is at least that of the modal systems KD and KT, respectively, in the Chellas classification [3].

First take the four positive expressions BKp, $BK\neg p$, $B\neg Kp$, and $B\neg K\neg p$, and then the corresponding negative expressions $\neg BKp$, $\neg BK\neg p$, $\neg B\neg Kp$, and $\neg B\neg K\neg p$. These eight expressions can obviously be arranged as four truth-functional tautologies, as follows:

$$
\begin{array}{ll}
T1. & BKp \lor \neg BKp \\
T2. & BK\neg p \lor \neg BK\neg p \\
T3. & B\neg Kp \lor \neg B\neg Kp \\
T4. & B\neg K\neg p \lor \neg B\neg K\neg p
\end{array}
$$

There are 16 ways of selecting precisely one disjunct from each of $T1.$ $-T4.$, to form a conjunction of four conjuncts. Of these 16 conjunctions just 6 are logically consistent, given the logical properties adopted for the two modal operators. These are listed by Pörn (op. cit., p.208) as the 'atomic epistemic positions':

$$
\begin{array}{ll}
EP1. & BKp \land \neg B\neg Kp \land \neg BK\neg p \land B\neg K\neg p \\
EP2. & \neg BKp \land B\neg Kp \land BK\neg p \land \neg B\neg K\neg p \\
EP3. & \neg BKp \land B\neg Kp \land \neg BK\neg p \land B\neg K\neg p \\
EP4. & \neg BKp \land B\neg Kp \land \neg BK\neg p \land \neg B\neg K\neg p \\
EP5. & \neg BKp \land \neg B\neg Kp \land \neg BK\neg p \land B\neg K\neg p \\
EP6. & \neg BKp \land \neg B\neg Kp \land \neg BK\neg p \land \neg B\neg K\neg p
\end{array}
$$

These 6 positions are mutually exclusive, and their disjunction is a logical truth. Thus, precisely one of $EP1.$ – $EP6.$ must hold for any given proposition p.

Pörn describes EP6. as the epistemic null-position, and doubts whether it is the characteristic epistemic position of any emotion, "...since in this position the subject has no belief at all concerning p" (op. cit., p.208). (However, he explicitly notes that the epistemic null-position might itself be the object of some emotion; we might for instance imagine a situation in which someone regrets having no beliefs about p, in which case he would have an epistemic position of type certainty with respect to the fact that he has no beliefs about p.)

Each of the remaining 5 conjunctions may be simplified by removing those conjuncts that are themselves logically implied by one or more other conjunct. The result is the following list of simplified atomic epistemic positions:

$SEP1.$ BKp

$SEP2.$ $BK\neg p$

$SEP3.$ $B\neg Kp \ \wedge \ B\neg K\neg p$

$SEP4.$ $B\neg Kp \ \wedge \ \neg BK\neg p \ \wedge \ \neg B\neg K\neg p$

$SEP5.$ $B\neg K\neg p \ \wedge \ \neg BKp \ \wedge \ \neg B\neg Kp$

2.2 The Volitional Component

In his paper, Pörn does not offer a modal-logical representation of the volitional component, but it seems clear from what he does say that he has in mind a desire/wish propositional modal operator of type KD, since he identifies three distinct volitional positions that correspond exactly to the three elementary normative positions of Standard Deontic Logic (which is also a modal-logical system of type KD – see [2]). These three positions are, respectively: a desire that p, a desire that $\neg p$, and the 'indifference' position in which there is neither a desire that p nor a desire that $\neg p$. Introducing an operator, call it D, we get:

$VOL1.$ Dp

$VOL2.$ $D\neg p$

$VOL3.$ $\neg Dp \ \wedge \ \neg D\neg p$

As for the epistemic null-position, Pörn doubts the relevance of the volitional indifference position to the analysis of emotions. "Do we not require of every emotion that it exhibits a genuine will – in addition to a non-empty epistemic attitude?" (op. cit., p.208). (However, he would no doubt agree that indifference might itself be the object of some emotion: consider, as a possible candidate, being ashamed of one's indifference.)

Pörn next proceeds to bring together the 5 epistemic and 2 volitional components to generate a set of 10 atomic emotional types, which we may represent by conjoining Dp and $D\neg p$, respectively, to each of $SEP1. - SEP5.$ As before, we omit the subscripts to the modal operators and assume, in the following list, that the epistemic and volitional positions are those of the same individual:

$EM1.$ $BKp \ \wedge \ Dp$

$EM2.$ $BKp \ \wedge \ D\neg p$

$EM3.$ $BK\neg p \ \wedge \ Dp$

$EM4.$ $BK\neg p \ \wedge \ D\neg p$

$EM5.$ $B\neg Kp \ \wedge \ B\neg K\neg p \ \wedge \ Dp$

$EM6.$ $\quad B\neg Kp \ \wedge \ B\neg K\neg p \ \wedge \ D\neg p$

$EM7.$ $\quad B\neg Kp \ \wedge \ \neg BK\neg p \ \wedge \ \neg B\neg K\neg p \ \wedge \ Dp$

$EM8.$ $\quad B\neg Kp \ \wedge \ \neg BK\neg p \ \wedge \ \neg B\neg K\neg p \ \wedge \ D\neg p$

$EM9.$ $\quad B\neg K\neg p \ \wedge \ \neg BKp \ \wedge \ \neg B\neg Kp \ \wedge \ Dp$

$EM10.$ $\quad B\neg K\neg p \ \wedge \ \neg BKp \ \wedge \ \neg B\neg Kp \ \wedge \ D\neg p$

The emotional types represented by $EM2.$ and $EM3.$ are both species of *regret* (Pörn calls them *despair*) since, in both cases, x is certain that what is the case is the opposite of that which he desires. By contrast, $EM1.$ and $EM4.$ both represent situations in which what x takes to be certain matches what he desires. Pörn used the term *security* in connection with these two types, and we shall return to them below in a discussion of *trust*. $EM5.$ and $EM6.$, represent *anxiety*, since x believes that, for all he knows, p might be the case and that, for all he knows, p might not be the case – and in the one case he desires p, in the other $\neg p$. Positions $EM8.$ and $EM9.$ may be understood to represent *hope*: although x is not certain that that which he desires is the case, he nevertheless believes that the realisation of his desire is compatible with all that he knows. Parallel considerations would lead one to interpret positions $EM7.$ and $EM10.$ as species of *fear*.

2.3 Complex Emotions: Trust and Shame

Pörn discusses a number of ways in which these atomic types can provide a basis for articulating the structure of other emotions. As hinted above, one of these ways involves restricting the object of the emotion, represented by p in the formulae above, to particular types of states of affairs.

Let us try to illustrate the applicability of this idea by considering the central proposal in [4] concerning the characterisation of *trust*. Jones proposed that the object or content of an *attitude of trust* consisted of two elements: first, that there existed a rule – for instance, a norm placing some agent under an obligation; and secondly that the rule would be complied with – for the case just mentioned, the compliance would come in the form of the fulfilment of the agent's obligation. In Jones' account the truster's attitude was represented simply as belief, rather than certainty of the type expressed by BK; accordingly, the truster was characterised in terms of his rule-belief and his conformity-belief. Now, suppose that in $SEP1.$ – $SEP5.$, above, the proposition p describes the state of affairs that there exists a rule and that that rule is complied with. Three of positions $SEP1.$–$SEP5.$ are compatible with the truth of Bp: $SEP1.$ logically implies Bp, because the belief operator is closed under logical consequence and Kp logically implies p; Bp can also be consistently added as a conjunct to $SEP3.$, and if it is the resulting conjunction can be simplified to $Bp \ \wedge \ B\neg Kp$; finally, Bp can be consistently added as a conjunct to $SEP5.$, and if it is the resulting conjunction can again be simplified, this time to $Bp \ \wedge \ \neg BKp \ \wedge \ \neg B\neg Kp$. So then, these considerations would suggest three mutually exclusive candidates for the epistemic/doxastic component of trust, where the proposition p is given the specific, restricted interpretation indicated above:

$E/DTRUST1.$	BKp
$E/DTRUST2.$	$Bp \wedge B\neg Kp$
$E/DTRUST3.$	$Bp \wedge \neg BKp \wedge \neg B\neg Kp$

Is trust an emotion? If so, on Pörn's view it cannot be characterised in terms of an epistemic component alone. So let us now turn attention to the question of whether trust might also be said to contain a volitional aspect.

In section 5 of [4] it was noted that one could make perfectly good sense of the idea that an agent x had beliefs to the effect that a rule held, and would be complied with, even though x did not care about whether there would be compliance with the rule, and perhaps remained indifferent to the existence of the rule itself. For instance, x believes that Norwegian bureaucrats are under an obligation to collect subsidies from cod fishermen, and believes that they will fulfil that obligation, but has no desires either way about this object of his trusting attitude. It was also noted, however, that this is a somewhat eccentric scenario, in as much as the contexts in which we are primarily concerned with matters of trust are those in which compliance matters to the truster; x perhaps stands to lose if compliance is not forthcoming, and it is in part because we most commonly associate trust with situations in which the truster is not indifferent that we tend to link trust closely with risk. One is inclined to say, in the spirit of Pörn, that when trusting beliefs are combined with indifference, then the truster lacks involvement, lacks engagement... there is no emotional edge to his trust. So then we might conjoin each of *E/DTRUST1.* – $EDTRUST3.$ with Dp, to arrive at three candidates for representing trust as a type of emotion:

$EMTRUST1.$	$BKp \wedge Dp$
$EMTRUST2.$	$Bp \wedge B\neg Kp \wedge Dp$
$EMTRUST3.$	$Bp \wedge \neg BKp \wedge \neg B\neg Kp \wedge Dp$

(Incidentally, were we to conjoin them instead with $D\neg p$, then we would have structures close, it seems, to Castelfranchi and Falcone's notion of 'aversive trust' – see section 4 of [5].)

It is interesting to observe here that this way of viewing *trust* helps to place it more clearly in relation to its near neighbour *hope*. For while it may well be agreed that $EMTRUST1.$ does fit intuitively with the concept of *trust* (remember that the proposition p has been given a specific interpretation concerning rule-existence and compliance, as indicated above), it might well be suggested that $EMTRUST2.$, given the uncertainty expressed by its second conjunct, is more akin to hope, with $EMTRUST3.$ perhaps exhibiting a 'strength' that falls somewhere between *trust* and *hope*.

In our opinion, what we have here is a good example of the analytical value of these formal tools, which perhaps also brings out the futility of trying to 'force' the vague notion of trust into one particular mould. The analytical tools enable us to articulate the spectrum of concepts to which phenomena of type *trust* belong. No single point on that spectrum tells the whole story about trust. But when we have a clear, preferably formal-logical model of that spectrum we can, in designing particular systems for particular applications, identify the points on the spectrum of most relevance to the requirements specifying the task at hand. (We say "preferably formal-logical model" because of the obvious advantages such models bring in terms of testing for consistency and for relations of implication.)

As a second illustration of the expressive power of this formal framework, we turn attention to aspects of the concept of *being ashamed*. Pörn indicated that a second way in which other emotional types could be generated from the atomic types was by taking emotions as themselves the objects of emotions. We illustrate this possibility by offering an analysis of one interpretation of *shame*. (We are not here suggesting that Pörn himself would choose to analyse *shame* in just the same way.)

The key idea is to understand shame with respect to one's action, or failure to act, as regret that some other party regrets that action/failure to act. Suppose, first, that y regrets that x has not done q. The object of the emotion may be represented by $\neg\mathcal{E}_x q$, where $\mathcal{E}_x$ is a relativised modal action operator of, for instance, the type employed and defined in [2]. y's regret that x has not done q may be understood as y's certainty that x has not done q coupled with y's desire that x has done q:

$$(REG1) \qquad\qquad B_y K_y \neg\mathcal{E}_x q \;\wedge\; D_y \mathcal{E}_x q$$

We now suggest that x is ashamed, vis-à-vis y, that he (x) has not done q if ($REG1$) is itself the object of an emotional state of type *regret* on the part of x:

$$(ASH1) \qquad B_x K_x (B_y K_y \neg\mathcal{E}_x q \;\wedge\; D_y \mathcal{E}_x q) \;\wedge\; D_x \neg(B_y K_y \neg\mathcal{E}_x q \;\wedge\; D_y \mathcal{E}_x q)$$

A further instance of this type of shame would be one in which it is not only the case that x fails to do q, but also the case that x was under an obligation to do q: i.e., x's shame vis-à-vis y that he (x) has violated an obligation. We could formally articulate this type of case if we further enrich the modal-logical language by introducing a *directive* normative modality – in contrast to an *evaluative* normative modality, of which the *desire* modality is an example. (On the distinction between evaluative and directive normative modalities see [6]. Pörn attributed articulation of this distinction to Kanger.)

3 Points of Departure

In this section, we indicate how the formal characterisation of the previous section offers a common platform for three points of departure for further work. Firstly, we consider an investigation of how this modal-logical taxonomy of emotional types compares to psychophysiological taxonomies. Secondly, we consider the relation between the modal-logical taxonomy, the understanding of trust it offers, and its relation to the concept of forgiveness. Thirdly, we consider an investigation into how this systematic, modal-logical characterisation of emotional types might provide a 'middle-layer' for enhancing previous work in principled operationalisation of the classification, to support multi-agent systems engineering.

To begin with, though, we briefly articulate our methodology to clarify the role of this kind of formal analysis and its contribution to specifying and implementing cognitive agent architectures.

3.1 Cognitive Agent Architectures

Methodologically, a standard practice in intelligent agents research and agent-based social simulation is to study a formal theory, formalise it in some logic, process or architecture, and then operationalise it in a computational implementation. This process has been apparent in much research aiming to develop cognitive agent architectures, i.e. a precise specification and implementation of a software agent capable of performing cognitive reasoning with socio-cognitive concepts and relations.

Relevant examples include: (1) the study of Dennett's intentional stance [7], its formalisation in modal logic [8], and its operationalisation in agent programming environments such as Jason [9]; (2) the study of Searle's speech act theory [10], its formal characterisation in terms of mental states (e.g. FIPA-ACL), and its operationalisation in (supposedly) FIPA-compliant agents; and (3) the study of Searle's speech act theory (op. cit.), its formal characterisation as institutional power, 'counts-as' and constitutive norms [11], and its operationalisation in different action languages [12].

Our point here is to emphasise the role of the formal characterisation, which is three-fold. The first is to bring conceptual clarity and precision to the underlying theory, which is often couched in natural language; the second is to provide sufficient abstraction and a 'toolbox' to explore 'joins' between theories; and the third is to provide a platform for the operationalisation. Therefore the formal-logical characterisation is not necessarily expected to provide direct computational support, but it is expected to be a calculus of some sort, i.e. any system of calculation or computation based on symbolic representation and manipulation.

One can see by the formal 'moves' made in the previous section that the modal-logical language is fulfilling this calculus requirement. In the following sub-sections, we indicate its role in relation to psycho-physical taxonomies of emotion, psychological theories of forgiveness, and agent architectures.

3.2 Psychophysiological Classifications of Emotion

It has been well-recognised that emotions play a critical role in cognitive processes in humans: therefore there is interest in being able to identify and classify affective states. For example, a cross-cultural study [13] of facial recognition identified six basic emotions: anger, fear, disgust, surprise, happiness and sadness (see Figure 1).

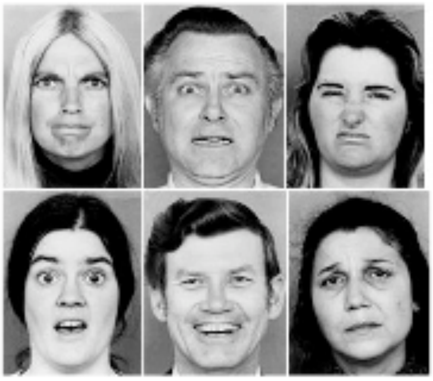


Fig. 1. Emotions displayed by facial recognition

Subsequently one of the most influential theories of emotion is the psychoevolutionary model of Plutchik [14]. Plutchik argued that emotions offered an evolutionary advantage in either reproduction or survival, and held that there were eight of these in four bi-polar pairs: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation. Figure 2 represents Plutchik's conical model, indicating the pairs, the different intensities, and mixtures which produce different emotions.

One of the ideas of affective computing [15] is that physiological signals are an indicator of psychological (emotive) state; and that signal processing techniques can be used, in conjunction with a theory of emotion, to infer that emotional state from recorded signals. Experiments have shown that it is possible to identify a change of a physiological measure (galvanic skin response) associated with a change of affective state. Therefore it is possible to ground a natural language label with both a facial display and a physical signal. This suggests that emotions have a definite ontological status (unlike, say, phonemes, which were revealed not to be associated with an event in an acoustic waveform) and can therefore be categorised with a formal representation.

We conjecture here that it is possible to ground the psycho-physiological taxonomy of Plutchik in the formal characterisation of Pörn, though this is a matter for further research. However, we can make the following two preliminary observations.

Firstly, it is useful to observe where the psychophysiological taxonomy of Plutchik and the modal-logical taxonomy of Pörn do, and do not align. For example, the atomic emotional types of Pörn were *security*, *despair* (*regret*), *anxiety*, *hope* and *fear*. On the grounds that 'the world' is, and is not, the way the agent wants it, *security* and *despair* seem to correspond to joy and sadness; *hope* to anticipation, and *fear* to fear. The two flavours of *anxiety* appear to

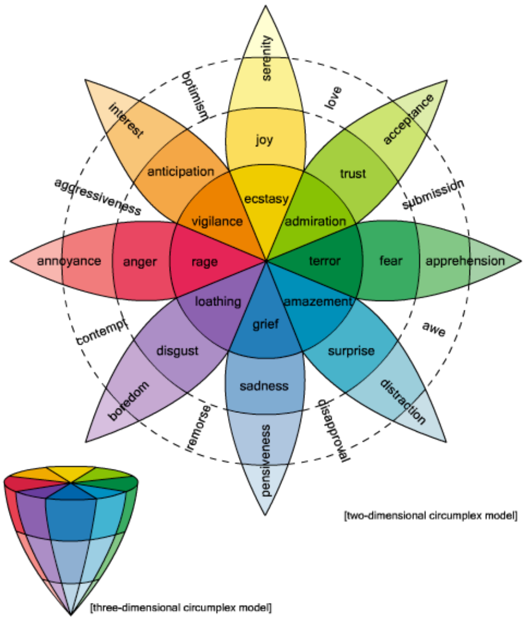


Fig. 2. Plutchik's Conic Model of Emotion

correspond to interest[1] and apprehension, i.e. the weaker (in intensity) versions of Plutchik's anticipation and fear (which leads to the speculation that the emotive intensity can be correlated with the epistemological strength, but this is for further work). Finally, the atomic emotions of Plutchik, trust and disgust, seem to correspond to *complex* emotion types of Pörn, as constructed here, namely $EMTRUST1. - EMTRUST3.$, and $ASH - REG1$, respectively.

Secondly, the key intuition of Pörn, that there are some emotions that are compatible with certainty, and some which are not, is consistent with the bi-polar distinctions of Plutchik, in that surprise and anger require certainty, while fear, and anticipation are incompatible with it. However, joy and sadness both seem to require certainty; while the other bi-polar pair, trust and disgust, exhibit a range of positions both compatible and incompatible with certainty.

We believe that there is rich ground to be explored here, but the key point we wish to emphasise in this workshop paper is that an exercise of this sort can be beneficial in both directions, in that it may lead to re-appraisal of either the psychophysiological theory, or the formal characterisation.

[1] A better term would be engagement: the agent is concerned with or has an interest in some p, but not unduly bothered.

3.3 Forgiveness

To characterise an *attitude* of trust, as we suggested above, the trustor should hold hold two beliefs [4]: that there is a rule, and that someone else's behaviour will conform to that rule. To characterise a *decision* to trust, the trustor should make a computation: what is the probability that someone's behaviour will conform to that rule, and what is benefit/cost if someone's behaviour does/does not conform to that rule [16]. This varies according to context, and so the trust decision ranges from *strong* trust, as characterised by $EMTRUST1.$, which is a kind of a short-cut to offset the more 'expensive' computation required for the *weaker* trust assessmenta as characterised by $EMTRUST2. - EMTRUST3$.

For there to be a any kind of trust decision, there has to be (objectively) a finite probability of error: if As a result, in either strong or weak trust, there is always a possibility that the decision may be wrong; and an essential element of trust, often overlooked, is what to do when the trust decision goes wrong.

In [17, 18], a forgiveness mechanism was proposed for decision-making about violation of norms. This was not based on reputation, which is a quantitative punishment mechanism, but instead on forgiveness. This is a qualitative repair mechanism, known from psychological studies to stimulate voluntary acts of recompense, reduce a negative predisposition towards an offender, and accentuate a positive motivation for self-repair.

The forgiveness framework defined in [17] comprises eleven constituent signals (severity, frequency and intent of the offence; apology or reparation; utility and frequency of beneficial relationship; and familiarity, similarity, and shame or embarrassment) underlying the four positive motivations relating to the nature of the offence, remedial action, historical record and empathic relationship. This was implemented using a fuzzy inference system (FIS) which used fuzzy rules to compute a fuzzy value for each of the four positive motivations from the respective signals, which were themselves combined by a fifth FIS to output a forgiveness decision (see Fig. 3; note that equal weight (1/4, or 25%), is given to each positive motivation, but different relative strengths can also be used).

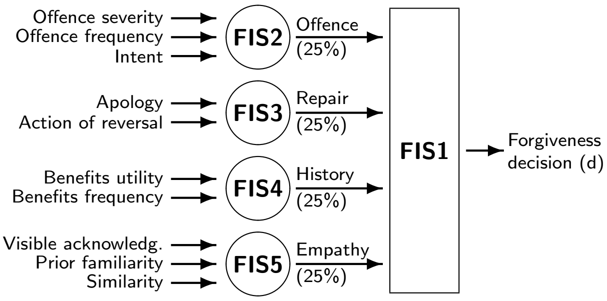


Fig. 3. Forgiveness framework

We can now see how the nuanced analysis of a 'spectrum' of trust attitudes $EMTRUST1. - EMTRUST3.$ could be inputs for the forgiveness decision. For example, for the offending party, the 'strength' of the trust would correlate with the perceived seriousness of the breach (of trust). That, taken together with any feelings of regret ($REG1$), based on one's self-evaluation, plus a feeling of shame (ASH), based on one's self-evaluation of oneself as seen by others, may be what triggers an apologies or other actions of repair, as well as involuntary reflex actions such as blushing. These are inputs to the forgiveness model of the offended party (apology or action of reversal signals for FIS3, and visible acknowledgment signal for FIS5), while the similarity signal could use the same modal-logical reasoning ("well, if it were me..."). This correlation illustrates a point emphasised in [19], who observes that emotions contain an element of cognitive appraisal which drives a system towards a homeostatic equilibrium (where 'system' in this case, comprises two independent but inter-connected components, the offending and the offended parties, and the equilibrium is measured by the 'strength' of their emotions).

3.4 Principled Operationalisation

In section 3.2, we 'looked back', as it were, to the relation of the formal-logical characterisation to a psychophysiological theory; in the previous section, we looked at how the formal characterisation related to other theories at the same level of abstraction, i.e. a formal-logical characterisation (in fuzzy logic) of a psychological theory of forgiveness. The final direction is to 'look forward', towards implementation in intelligent agents and multi-agent systems. At this point, though, we can only mention a programme or work in which we also intend to relate the formal taxonomy to the approach of Digital Blush [20], which tried to encapsulate the cognitive appraisal identified by Castelfranchi [21] in an operational specification of action and agency; and the anticipatory agent architecture of [22] which integrates a model of intention-prediction from robotics [23] with affective appraisal of expectations [19].

We also mention this point because we see this programme of work as a further illustration of the importance of a distinct methodological strategy based on Steels' Synthetic Method [24, 16] – i.e. understanding a clear separation of theory, formal characterisation and principled operationalisation, particularly in the context of inter-disciplinary research of this kind.

We also offer the conjecture that the systematic formal characterisation of emotional types will provide a richer and more stable platform for characterising the spectrum of trust and forgiveness decisions than the constrained specification language of [25], and for the development of cognitive agent architectures than the types of formalism employed in, for example, [26] and [27].

4 Conclusions

This paper has offered a preliminary report on work in progress, in which we have outlined some principal features of Pörns modal-logical characterisation of some types of emotions, and indicated how that highly systematic approach might be applied to the analysis of trust and shame. We also identified some points of departure for further work: a detailed and systematic comparison of the formal-logical taxonomy of emotions with psychophysiological taxonomies; investigating responses, such as forgiveness, to the violation of trust and the 'feeling' of shame; and a possible development plan for an affective/cognitive agent architecture.

Much remains to be done, including development and discussion of the following four topics. Firstly, we have given only an initial indication of how the formal-logical theory, combining epistemic/doxastic and volitional modalities, might be applied to the representation of types of emotions. In particular, much remains to be said about the characterisation of such complex emotions as shame, embarrassment and regret, which exhibit subtleties of a kind that call for further discussion. This will in turn raise the question of whether the modal-logical language has sufficient expressive capacity as it stands; for instance, perhaps shame requires incorporation of the notion of obligation, since in some instances at least shame appears to involve a belief that one has violated an internalised norm.

Secondly, we should explore the question of whether it is appropriate to represent the desire modality in terms of a normal modality, since perhaps use of a normal modal logic will prevent adequate analysis of conflicts of desires. Similarly as has often been observed the use of normal modalities for knowledge and belief requires an agent to know/believe all of the logical consequences of what it knows/believes. A switch to classical modalities and minimal-model semantics might therefore be deemed necessary. We note that a change of this sort would not require a revision of the method for generating the relevant classes of modal positions, although of course the resulting classes will be different, depending as they do on the specific logical properties assigned to the component modalities.

Thirdly, we frequently wish to talk of, for instance, levels of trust, and of degrees of shame or embarrassment. Here we should explore the application of graded modalities to capture the idea of strength of belief (degree of certainty), and the notion of level of desire [28].

Fourthly and finally, it has been brought to our attention that some very recent formal-logical work has interpreted such emotions as regret and disappointment as counterfactual emotions. That work may well provide a further line of development of the framework we have proposed in this paper [29, 30].

Acknowledgements

We are most grateful to the three reviewers for their constructive comments and for pointing us in the direction of a range of issues that remain to be addressed. In fact, the discussion in Section 4, above, relates to topics identified in one or more of the three reviews. This advice is very much appreciated.

References

1. Pörn, I.: On the nature of emotions. In: Changing Positions, Philosophical Studies. Volume 38. Department of Philosophy, University of Uppsala (1986) 205–214
2. Jones, A., Sergot, M.: Deontic logic in the representation of the law: Towards a methodology. Artificial Intelligence and Law **1**(45–64) (1992)
3. Chellas, B.: Modal Logic – an introduction. CUP (1980)
4. Jones, A.: On the concept of trust. Decision Support Systems **33**(3) (2002) 225–232
5. Castelfranchi, C., Falcone, R.: Social trust: a cognitive approach. In Castelfranchi, C., Tan, Y.H., eds.: Trust and Deception in Virtual Societies. Kluwer (2001) 55–90
6. Pörn, I.: Action Theory and Social Science – Some Formal Models. Volume 120 of Synthese Library. Reidel (1977)
7. Dennett, D.: The Intentional Stance. MIT Press (1987)
8. Cohen, P., Levesque, H.: Intention is choice with commitment. Artificial Intelligence **42** (1990) 213–261
9. Bordini, R., Hübner, J., Wooldridge., M.: Programming Multi-agent Systems in AgentSpeak Using Jason. Wiley (2007)
10. Searle, J.: Speech Acts: An Essay in the Philosophy of Language. CUP (1969)
11. Jones, A., Sergot, M.: A formal characterisation of institutionalised power. Journal of the IGPL **4**(3) (1996) 429–445
12. Artikis, A., Sergot, M., Pitt, J.: Specifying norm-governed computational societies. ACM Transactions on Computational Logic **10**(1) (2009)
13. Ekman, P., Friesen, W.: Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. Prentice-Hall (1975)
14. Plutchik, R.: Emotion: A psychoevolutionary synthesis. Harper & Row (1980)
15. Picard, R.: Affective Computing. MIT Press (1997)
16. Neville, B., Pitt, J.: A computational framework for social agents in agent mediated e-commerce. In Omicini, A., Petta, P., Pitt, J., eds.: Engineering Societies in the Agents World (ESAW) IV. Volume 3071 of LNCS., Springer (2004) 376–391
17. Vasalou, A., Pitt, J., Piolle, G.: From theory to practice: Forgiveness as a mechanism to repair conflicts in cmc. In 397-411, ed.: Proceedings iTrust IV. (2006)
18. Vasalou, A., Hopfensitz, A., Pitt, J.: In praise of forgiveness: Ways for repairing trust breakdowns in one-off online interactions. International Journal of Human-Computer Studies **66** (2008) 466–480
19. Castelfranchi, C., Lorini, E.: Cognitive anatomy and functions of expectations. In: IJCAI'03 Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions. (2003) 9–16
20. Pitt, J.: Digital blush: towards shame and embarrassment in multi-agent information trading applications. Cognition, Technology and Work **6**(1) (2004) 23–36
21. Castelfranchi, C.: Affective appraisal vs cognitive evalutation in social emotions and interactions. In Paiva, A., ed.: Affective Interactions. Towards a New Generation of computer Interfaces. Springer (2000) 76–106
22. Sanderson, D., Pitt, J.: Anticipatory agent architecture for defensive driving. (forthcoming)
23. Demiris, Y.: Prediction of intent in robotics and multi-agent systems. Cognitive Processing **8**(3) (2007) 151–158
24. Steels, L., Brooks, R.: The artificial life route to artificial intelligence: Building Situated Embodied Agents. New Haven: Lawrence Erlbaum Ass (1994)
25. Grandison, T., Sloman, M.: Specifying and analysing trust for internet applications. In: 2nd IFIP Conference on e-Commerce, e-Business, e-Government. (2002)

26. Meyer, J.J.C.: Reasoning about emotional agents. In: Proceedings of the 16th Eureopean Conference on Artificial Intelligence. (2004) 129–133
27. Ochs, M., Sadek, D., Pelachaud, C.: A formal model of emotions for an empathic rational dialog agent. Autonomous Agents and Multi-Agent Systems (2010)
28. Lorini, E., Demolombe, R.: From binary trust to graded trust in information sources: A logical perspective. In: AAMAS-Trust. Number 5396 in LNCS, Springer (2008) 205–225
29. Bonnefon, J.F., Longin, D., Nguyen, M.H.: Relation of trust and social emotions: A logical approach. In: Proceedings of IAT. (2009) 289–292
30. Lorini, E., Schwarzentruber, F.: A logic for reasoning about counterfactual emotions. Artificial Intelligence **175**(3-4) (2011) 814–847

A Generative Foundation for Trust Networks

H. Van Dyke Parunak

Vector Research Center, a business unit of Jacobs Technology
3520 Green Court, Suite 250
Ann Arbor, MI 48105
van.parunak@jacobs.com

Abstract. Representation of trust among agents as a network has become *de jure* in formal models of trust. Often, these networks are compiled directly for purposes of assessing trust, or assumed to be identical to a social network underlying the community. We show how a trust network can be generated from the graph union of three underlying structures: a social network of who knows about whom, a hierarchical task network describing the actions that define the scope of trust, and an assignment network linking the other two. This approach provides additional insight into key characteristics of trust relations, including scope, type, and process. While the graph union can become extremely large and difficult to search, swarming methods provide a tractable way to explore it for crucial features of the trust network that it generates.

Keywords: trust network, social network, task network, HTN, generative models

1 Introduction

The literature on agent trust commonly represents trust as a network [11, 17], a directed graph $G_f = \langle A, E_f \rangle$ where $A = \{a_i\}$ is a set of agents and $E_f \subset A \times A$; an edge from a_i to a_j means that a_i has an estimate of its trust in a_j. We say that there is a *trust relation* between a_i and a_j.[1] The structure of G_f is often taken as given, and the edges are adorned with a variety of features. One convenient ontology [2] includes

- The **scope** of action for which the trustee is being trusted;
- The **type** of trust, usually distinguishing functional trust (to do something) from referral trust (to recommend someone else to do something) [2, 15, 17, 31];[2]

[1] The existence of a trust relation between Alice and Bob does *not* mean that "Alice trusts B." Alice might in fact *dis*tust Bob. The level of trust is a characteristic of the trust relation, as discussed below, but the relation itself simply means that Alice and Bob stand in a relation in which it is meaningful to talk about whether or to what degree Alice trusts B. Alternatively, using Jøsang's model of opinion space [16], what we call a trust relation is a link in a social network along which the uncertainty is less than 1.

[2] The f in G_f is for functional trust. The analogous referral trust network is $G_r = \langle A, E_r \rangle$. In Section 3.3, we will suggest that this distinction may be overdrawn.

- Some measure of the **level** of trust; and
- The **process** that generates the trust.

The thesis of this paper is that

- G_f and G_r are not primitive, but generated from other more fundamental graph structures in the domain;
- This perspective can refine and sharpen our notions of just what trust is;
- Attending to these more fundamental structures and how they generate G_r can give useful insight into the characteristics associated with edges in E_f and E_r.

Section 2 identifies the primitive structures that generate of G_f and G_r, and defines G_f in terms of these structures. Section 3 shows how this definition can help characterize individual trust relations. Section 4 discusses how to deal with the complexity that this construction appears to impose. Section 5 concludes.

2 Generating Trust Networks

Trust networks are generated by the interplay of three more fundamental graphs: a *social network,* a *task network*, and an *assignment network*. We explore each of these in turn, and then consider how together they generate G_f.

2.1 Three Fundamental Graphs

A *social network* is a directed graph $G_a = \langle A, E_a \rangle$ where $A = \{a_i\}$ is a set of agents and $E_a \subset A \times A$ in which an edge from a_i to a_j means that a_i knows about a_j. We do not require E_a to be symmetric. The agents in G_f are also in G_a, and one can view the trust network as the social network with additional adornments on the edges. In fact, our G_f will be a subgraph of G_a; a_i cannot trust a_j if a_i does not know that a_j exists. But one insight from our generative definition is that acquaintances in G_a need not participate in a trust relation. By distinguishing the underlying social fabric from the trust relationships that it generates, the characteristics of trust become less arbitrary.

A *task network* is a directed acyclic graph $G_t = \langle T, E_t \rangle$ where $T = \{t_i\}$ is a set of tasks and $E_t \subset T \times T$ in which an edge from t_i to t_j means that t_j is a subtask of t_i. G_t is commonly called a hierarchical task network, or HTN [7]. More can be said about such a network [14, 20], including whether completion of a task requires that all or only some of its subtasks be completed and how the task sequence is constrained. Such enhancements can be applied to the generative program, but the simple HTN suffices to illustrate the approach.[3] In engineered systems such as sensor networks or communication protocols, G_t often emerges directly from the system design.

[3] Our construction generates trust relations only through subtask links. Sequence constraints can also generate trust relations: if Alice is responsible for a task, and Bob is responsible for something that must be completed before Alice's task can start, Bob's task is not a subtask of Alice's, but it still makes sense to reason about the trust relation between Alice and Bob. Our approach is easily extended to include sequence-based trust relations.

An *assignment network* is a bigraph $G_s = \langle A \cup T, E_s \rangle$ where $E_s \subset A \times T$ in which an edge from a_i to t_j means that a_i can be assigned t_j. We should avoid two misunderstandings about an edge $<a_i, t_j>$. 1) It does not mean that a_i is able to do t_j. Trust is necessary because agents sometimes are *not* able to do all that they are expected to do by themselves, and must trust others to help them. 2) It does not mean that someone has assigned a_i to do t_j. Such an assignment might result from a trust relation, but before that relation can exist, the trustee must be thought relevant to the task for which she is trusted. One way to realize the assignment network is through the "external description" [6] that forms the basis for the widely used dependence network model of social interaction [27, 28]. An agent's external description consists of the information about it that may be known by other agents, and in the dependence network model includes the actions that the agent is able to perform.[4]

2.2 Generating the Functional Trust Network from the Fundamental Graphs

Our generation of G_f formalizes a pregnant comment in [33]: "If Alice trusts Bob…, then this means that Alice is putting part of her plans in Bob's hands." Other intuitions are possible; our point is not that this is the best characterization of trust, but that the generative approach can capture this characterization, and so should be considered for other characterizations as well. We focus on functional trust [2] (trusting somebody to do something, rather than to recommend another agent), but our construction has a contribution to make to referral trust as well.

On this intuition, for Alice to have a trust relation with Bob (i.e., for the link from Alice to Bob to be in E_f), three things must be true, each corresponding to one of the foundational graphs:

1. Alice must know of Bob (accounted for by G_a);
2. Alice and Bob must each be associated with tasks (edges in G_s), otherwise there would be no reason for Alice to trust anyone and Bob would have nothing for which to be trusted;
3. Alice's task in G_s must have subtasks that she wishes to delegate[5] to someone else, and Bob's task in G_s must be one of those subtasks (accounted for by G_t).

Alice and Bob may not have the same perception of G_s and G_t. Our construction only requires that they share *local* views of these graphs. If they do not, our

[4] This information is subjective (an agent may in fact claim the ability to do actions that it cannot), but we reason about trust just because the engagement between two agents is uncertain [16].

[5] The vagueness of this condition recognizes that Alice's desire to delegate may stem from different reasons. For example: the subtasks may not be among her actions (possible if the subtask is ORed into the parent task, and the agent wishes to provide a backup for those subtasks that she can perform herself), or they may require resources that she does not possess (she could do the task, but just now she is out of time), or she may be able to do the tasks but wants to show her customer a team to persuade the customer that the effort is robust to failure of any one member. These various drivers are interesting, but beyond our scope in this paper.

construction explains how this difference in perception can modulate the characteristics of the trust relation between them.

To generate G_f from G_a, G_t, and G_s, we need two operators.

In G_s the *restriction of T to an agent a_i* is the set of tasks associated with a_i:

$$T|a_i \equiv \{t_j : \langle a_i, t_j \rangle \in E_s\} \qquad (1)$$

In G_t, the *subtasks* of a given task are

$$t_i^+ \equiv \{t_j : \langle t_i, t_j \rangle \in E_t\} \qquad (2)$$

In the sequel, it will also be useful to define the *restriction of A to a task t_i*,

$$A|t_i \equiv \{a_j : \langle a_j, t_i \rangle \in E_s\} \qquad (3)$$

and the *descendants* of a given task, which we define recursively:

$$t_i^* \equiv \{t_j : t_j \in t_i^+ \vee \exists t_k \in t_i^+ : t_j \in t_k^*\} \qquad (4)$$

We abuse notation to apply the subtask and descendant operators + and * to sets of tasks as well as individual tasks, in the natural way, e.g.,

$$(T|a_i)^+ \equiv \{t_j : \exists t_k \in T|a_i : t_j \in t_k^+\} \qquad (5)$$

We can now define G_f. Its nodes are just the set of agents A, and its edges are

$$E_f \equiv \{\langle a_i, a_j \rangle : \langle a_i, a_j \rangle \in E_a \wedge [(T|a_i)^+ \cap T|a_j \neq \emptyset]\} \qquad (6)$$

That is to say, Alice has a trust relation with Bob under the three conditions described informally above.

Fig. 1 shows this interplay of the three graphs pictorially. The dashed lines show the assignment network. The bold lines warrant the inference of the functional trust network *a* → *c* → *e*. Each link in this network corresponds to a closed undirected cycle of length 4 in the union of G_t, G_s, and G_a. The trust relation *a* → *c* corresponds to the cycle <*a, r, t, c, a*>, and *c* → *e* to <*c, t, v, e, c*>. There is no functional trust relation between *a* and *b* because of the structure of the task network: *b* is not assigned to a task that is a subtask of *a*'s task. There is no functional trust relation between *a* and *f* because of the structure of the social network: even though *f* is assigned to a subtask of *a*'s task, *a* does not know *f*.[6]

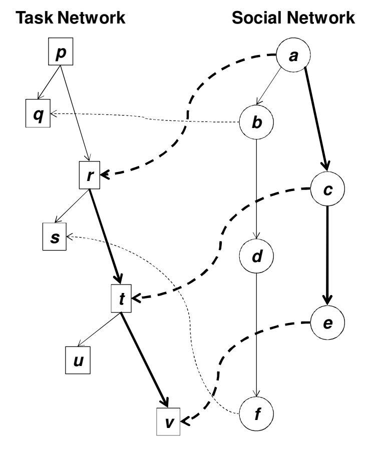


Fig. 1. Generation of functional trust network from task, social, and assignment networks

[6] One might ask whether the relation *a* → *c* makes sense if we delete the edge <*v, e*>. More generally, can *a* trust *c* to do *t* if *a* does not know in detail how *c* will do *t*? In the real world, this opacity of lower-level performers is the rule rather than the exception. Usually *a* does

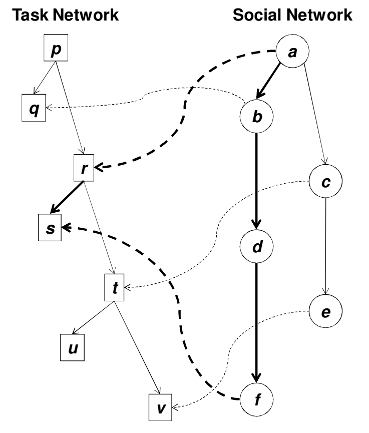


Fig. 2. Generation of referral trust network

2.3 Modeling Referral Trust

Fig. 1 also indicates how this model can contribute to the analysis of referral trust. *b* knows *d* and *d* knows *f*, so *b* is in a position to recommend *d* to *a*'s referral trust, and *d* can in turn recommend *f* to *a*'s functional trust. The bold lines in Fig. 2 show the relations that warrant the inclusion of $a \rightarrow b$ in E_r. A formal characterization of E_r along the lines of Eq. 6 is cumbersome because of the need to pass *a*'s assignment in the task network through arbitrary links in the social network in order to find the agent with whom *a* should establish a trust relation. However, Fig. 2 clearly shows that the structure that warrants a relation of referral trust, like that for functional trust, is an undirected cycle, this time of length > 4, in the union of G_t, G_s, and G_a. Identifying cycles in the graph union is thus a powerful heuristic for detecting trust relations of both types, a point to which we return in Section 4.

3 Characterizing Trust Relations

The justification for the previous section is the claim that deriving a trust network from more fundamental graphs can sharpen our sense of what trust is, and provide a rationale for some of the characteristics with which we want to adorn edges in E_f and E_r (again using the ontology in [2] as a convenient model).

3.1 What is Trust?

We motivated our model with a task-oriented view of trust [33]. This is not the only model of trust, but it is typical [8], and illustrates the value of decomposing the trust network into more primitive graphs. Any trust network clearly includes social relations, modeled in G_a. Once we suggest that Alice wants Bob to help her accomplish her plan, we introduce the notion of a plan, a construct with a long and

not have the time to chase the task tree all the way down; the distributed processing described in Section 4 can help. Sometimes *c* may not publish a list of those who help it do *t*, since that knowledge makes it competitive and if *a* knew of *e* directly, *a* might bypass *c*. Like the uncertainty in subjective assignments (footnote 4), this uncertainty highlights the need to frame the discussion in terms of trust. Far from invaliding our construction, it shows how the construction focuses our attention on possible source of uncertainty. If *a* knows the structure of the task network G_t, she knows the size of t^*, and thus can estimate her exposure to a failure of *c*'s trustees.

rich history in AI research. So we need to link our representation of trust to a formal model of a plan, for which a task network G_t is a well-known candidate.

More generally, any view of trust is a complex structure involving numerous cognitive components, each of which will likely be the object of serious study in its own right. By describing how a particular view of trust is generated from time-proven representations of each component, we can much more readily account for the interactions among them and draw on results that have been previously derived.

3.2 Scope

The scope of a trust relation is the domain in which the trustor is trusting the trustee to perform. The set of people whom I would trust to service my car has no members in common with the set whom I would trust to remove my appendix.

A trust system may be defined in a limited domain (such as a movie recommender system) with a single scope. In other cases, scope is considered an unstructured set of alternatives, or perhaps a hypercube of such alternatives, yielding a vector space in which different scopes may be defined [29].

For general applications, such approaches are unsatisfactory. Different scopes may be related to one another in ways that allow us to transfer trust from one scope to another. A major step forward in refining the notion of scope is the service graph proposed by Yolum and Singh [34], who construct a DAG of possible services for which one agent might trust another, with the semantics that if A trusts B for a higher-level service, A can surely trust B for a lower-level one.

In [34] the DAG is maintained by the trustor and exemplified by a sequence of services that are parameterized by an ordered scalar (the dollar value of the transaction that one trusts an agent to conduct). An HTN provides a more general way to relate different scopes to one another. Once we define the scope of a trust relation with reference to such a model, we can invoke the rich extensions to HTN theory [5, 14, 20] and powerful mechanisms for reasoning over them [3, 25] that are available as tools for reasoning about trust.

3.3 Type

Sometimes functional and referral trust are distinguished on the basis of transitivity [15]: referral trust is said to be transitive, while functional trust is considered non-transitive. Such a distinction is open to question, since making a recommendation can be viewed as just another task that one agent can entrust to another. Our use of an HTN to model scope gives more detailed insight into why this distinction should be qualified.

Functional trust is necessary in the first instance because tasks can be arranged in a subtask hierarchy, and because an agent assigned to a higher-level task seeks another agent to execute one of that task's subtasks. Once we recognize this hierarchical relationship, we must admit that the trustee's task may itself have subtasks, and that the trustee may invoke other agents (perhaps unknown to the original trustor) in addressing them. In this case, the trustor's trust in its immediate trustee has been transferred to that trustee's trustees, and so forth. Such derivative trust is again a cycle in the graph union, this time with multiple consecutive links in the task network (e.g., in Fig. 1, the outer cycle <*a, r, t, v, e, c, a*>).

This notion of derivative trust is important in many domains, such as manufacturing supply chains. Manufactured products are often composed of subassemblies, which themselves have smaller components. The top-level manufacturer (the "prime") contracts with one or a few companies (so-called "first tier suppliers") for the subassemblies, These companies in turn contract with lower, "sub-tier" suppliers for the lower-level components. In this case, the hierarchical product structure directly generates the HTN. The branching can be very high. In the case of automotive seats, major US auto manufacturers deal with about five companies (the ones who deliver finished seats to the automobile final assembly plant), but the network of sub-tier suppliers includes over 140 members, most of whom are not visible to the prime [1]. Primes are concerned about this lack of visibility. For example, the prime may hold contracts with multiple first-tier suppliers as back-up so that if one fails, another will be ready to provide the needed components. But if all of the first-tier suppliers depend on a single source for a critical component, this back-up capability is reduced. In terms of trust, manufacturing agents are very much aware that their functional trust in their immediate suppliers has a transitive component, and they would like to discover who their lower-level trustees are, a process known in industry as "achieving supply-chain visibility."

3.4 Level

In the introduction, we distinguished the existence of a trust relation between Alice and Bob from the question of whether Alice trusts or distrusts Bob. The latter question is addressed by the weight of the trust relation. A wide variety of representations have been considered, including natural numbers, real numbers in [-1, 1] (where negative numbers indicate distrust), or a partial ordering over a given trustor's trustees [31]. Recently, many researchers have adopted or built on Jøsang's representation of trust level as a triple <*a, b, c*> in opinion space, where a = degree of trust, b = degree of distrust, and $c = 1 - a - c$ = uncertainty [16]. This structure can be mapped directly to a couple <*r, s*> in evidence space, where r is the observed number of cases in which a trustee has satisfied trust and s the number of cases in which he has violated trust, but that mapping is a matter of trust process rather than level *per se*.

Our generative model does not tell us how to represent level of trust, though it does define some values that will be useful in quantifying the risk involved in a trust relation, such as the size of the set of alternative performers for a given task $A|t$ and the amount of the HTN t^* below a given trustee. In addition, it clarifies various proposals about the processes by which trust levels are computed.

3.5 Process

A "trust process" is the mechanism by which a trustor assigns a trust level to a trust relation. A number of such processes have been discussed. Most can be reduced to personal experience by the trustor of the trustee's past performance, and recommendations from other agents [10].

Our generative model suggests ways to modulate and extend both direct (past performance) and indirect (recommendation) evidence for the level appropriate to a given trust relation. Our interest here is not selecting among sophisticated proposals that have been made for how a given process updates trust values [12, 18, 21, 32, 35], but rather in how a generative view of trust networks enriches the inputs to any such algorithm.

Recognition of hierarchical structure in the scope of trust means that Alice can update her trust values for Bob based not just on her experience with subtasks of the specific task in her current scope, but also on his performance on related tasks. Consider the situation in Fig. 3, in which *a* has assignment *r*. Clearly, experience working with *b* assigned to *t* will increase her trust in *b*. However, *b*'s previous experience on *p* or *r* will also be relevant: if *b* is competent in performing a task that includes *t* as a subtask, it is likely to be able either to do *t* or to find some way to make *t* happen. (This reasoning draws on the transitivity of functional trust discussed in Section 3.3.) Alternatively, if *b* has done well *v*, a subtask of *t*, *a* may consider promoting *b* to *t* (using mechanisms such as those in [34]).

We have distinguished a social network defined by who knows of whom from a trust network (a subgraph of the social network). Paying attention to the social network as a first-class object in its own right can also extend our understanding of trust processes. For example, Falcone and Castelfranchi [8, 9] explore the role of trust in generating trust, including reciprocity effects. That is, there is likely to be a positive correlation between the level of trust from Alice to Bob and that from Bob to Alice. This kind of effect suggests an analogy to social balance theory [13, 19], an area that has yielded powerful quantitative tools [4, 30]. The underlying idea is that given a triangle in a social network (three people each of whom is connected to the other two), we can associate a valence (positive or negative) with each edge. The assignment of valences is stable if none is negative (all three people are friends), or two are negative (the two people with a positive valence agree in their negative attitude toward the third). However, the other two patterns of assignment are unstable. A single negative edge puts pressure on the person with two positive edges to side with one or the other end of the negative edge, while three negative edges motivates coalitions of two against one.

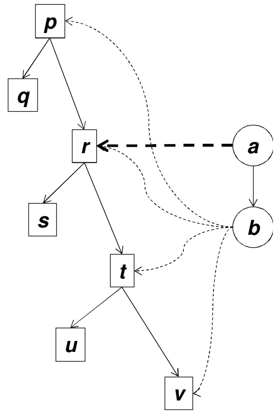


Fig. 3. Experience-based trust propagates through the scope HTN.

Falcone's insight suggests that we can extend this kind of reasoning to reputation theory. A pattern of trust values may be intrinsically unstable, for reasons analogous to those in social balance theory, completely unrelated to the actual evidence for an agent's trustworthiness. In addition, other sources of valence on the edges of the social network underlying a trust network may interact with the trust valence. Intuitively, one is more likely to trust someone toward whom one already has positive feelings. Understanding these dynamics takes us far beyond the scope of this paper, but we can raise the question only because we press behind the trust network to the more fundamental networks of which it is composed.

4 Reasoning over Generated Trust Networks

Generating trust networks from the underlying task, social, and assignment networks is mathematically elegant and offers new insights into trust relations, but it invites a pragmatic challenge. The graph union of a realistic task network, social network, and assignment network has two characteristics that can overwhelm conventional reasoning: it is very large, and it is distributed. That is, the global view of the network may not be available in one place. Each agent knows its acquaintances and has an estimate of its own abilities, but it may be technically or socially infeasible to centralize all this information. One solution to these challenges lies in highly parallel, decentralized processes.

In other domains, we have found swarm intelligence, a form of Monte Carlo modeling, a powerful tool for exploring such spaces. Our favorite technology, polyagents, uses multiple lightweight “ghosts” to search a complex space [26], coordinating their movements through probability fields modeled on insect pheromones [23], and generating such a field to record the likely outcome from sampling over multiple futures [24]. This technique can be applied to the graph union that generates the trust network. In particular, if we imagine that individual agents in the graph union maintain links to their neighbors, swarming ghosts can move from one agent to another, reasoning over the entire distributed network. Here are two examples.

Trust relations correspond to undirected cycles in the graph union of G_t, G_s, and G_a. Traditional functional trust corresponds to a cycle of length 4, referral trust and transitive functional trust to cycles longer than 4. Of course, not every cycle defines a trust relation. A cycle completely within the social network does not warrant a trust relation (though it may reflect a spurious one based on reputations running in a circle). Also, a cycle containing more than two edges from the assignment graph, or whose edges in that graph share an agent, is not useful.[7] In addition, an agent searching for trust partners is likely to prefer cycles whose edges in the social network have positive valence. Finding all the cycles in a large graph that meet such requirements is combinatorially explosive, but a swarming algorithm can easily identify the most likely candidates, which can then be verified in reasonable time.

More than one assignment of agents can execute a process represented as an HTN. Swarming algorithms can search an HTN for feasible partitioning of agents over tasks [3, 25], and polyagents can be trained from observed data, thus biasing our assignment of agents to tasks based on recent observed behavior [22]. These techniques allow us weight the edges in the assignment graph, and construct trust

[7] Note that the distinctions in the last two sentences cannot be made if the trust network is viewed as primitive, without decomposition into task, social, and assignment networks.

relations not just with any agent that has an association with a task, but with agents that are most likely to do a task, based on past experience. These algorithms can adapt in real-time, a particularly desirable attribute for a trust system supporting e-commerce processes.

5 Conclusion

Trust networks are a promising tool in managing modern distributed systems, but are not yet exploited as widely as they could be. Traditionally, they are generated directly from domain information. This paper argues that they can be derived instead from more fundamental graphs, and that viewing them as the product of a generative process offers useful insight into their characterization and suggestions as to how to manipulate them. This formalization may in turn expand their use and improve our ability to understand the systems on which we depend.

This proposal opens many directions for future work, which we look forward to exploring with our colleagues in the community. Here are some examples.

- The promise of the benefits of swarming over the graph union needs to be evaluated with actual experiments.
- Our suggestions about how the graph model informs trust processes are at this point qualitative. Numerous quantitative mechanisms have been proposed for computing trust values, and we need to integrate the expanded view of processes with those mechanisms. For example: How much weight should be given to experience on a task that is related to, but not identical with, the task that defines the scope of a present trust relation?
- The promise of an analog to, or merger with, structural balance theory for trust networks needs to be worked out in detail, drawing on solid psychological evidence for how valences of different types interact.

6 References

[1] AIAG. Manufacturing Assembly Pilot (MAP) Project Final Report. Automotive Industry Action Group, Southfield, MI, 1997. https://www.aiag.org/source/Orders/index.cfm?section=orders&activesection=AiagPubs&task=3&CATEGORY=MATM&PRODUCT_TYPE=SALES&SKU=M-4&DESCRIPTION=&FindSpec=&CFTOKEN=20008143&continue=1&SEARCH_TYPE=&StartRow=1&PageNum=1.

[2] P. Anantharam, C. A. Henson, K. Thirunarayan, and A. P. Sheth. Trust Model for Semantic Sensor and Social Networks: A Preliminary Report. In *Proceedings of National Aerospace & Electronics Conference (NAECON)*, 2010.

[3] S. Brueckner, T. Belding, R. Bisson, E. Downs, and H. V. D. Parunak. Swarming Polyagents Executing Hierarchical Task Networks. In *Proceedings of Third IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO 2009)*, pages 51-60, IEEE, 2009.

[4] D. Cartwright and F. Harary. Structural balance: a generalization of Heider's theory. *Psychological Review*, 63(5):277-293, 1956.

[5] K. Decker. *Environment Centered Analysis and Design of Coordination Mechanisms*. Thesis at University of Massachusetts, Department of Computer Science, 1995.

[6] Y. Demazeau and J.-P. Müller. From Reactive to Intentional Agents. In Y. Demazeau and J.-P. Müller, Editors, *Decentralized A.I. 2*, pages 3-10. Elsevier, Amsterdam, Netherlands, 1991.

[7] K. Erol, D. Nau, and J. Hendler. Semantics for Hierarchical Task-Network Planning. CS-TR-3239, UMIACS-TR-94-31, Computer Science Dept., University of Maryland, 1994.

[8] R. Falcone and C. Castelfranchi. The socio-cognitive dynamics of trust: does trust create trust? In F. R., S. M., and T. Y.H., Editors, *Trust in Cyber-Societies: Integrating the Human and Artificial Perspectives*, pages 55-72. Springer, 2001.

[9] R. Falcone and C. Castelfranchi. Trust Dynamics: How Trust is influenced by direct experiences and by Trust itself. In *Proceedings of the 3rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-04)*, pages 740-747, ACM, 2004.

[10] K. K. Fullam and K. S. Barber. Dynamically Learning Sources of Trust Information: Experience vs. Reputation. In *Proceedings of the International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'07)*, 2007.

[11] C.-W. Hang and M. P. Singh. Trust-Based Recommendation Based on Graph Similarity. In *Proceedings of the 13th AAMAS Workshop on Trust in Agent Societies*, 2010.

[12] C.-W. Hang, Y. Wang, and M. P. Singh. Operators for Propagating Trust and their Evaluation in Social Networks. In *Proceedings of the 8th International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2009.

[13] F. Heider. *The Psychology of Interpersonal Relation*. John Wiley & Sons, 1958.

[14] B. Horling, V. Lesser, R. Vincent, T. Wagner, A. Raja, S. Zhang, K. Decker, and A. Garvey. The Taems White Paper. Multi-Agent Systems Lab, University of Massachusetts, Amherst, MA, 2004. http://dis.cs.umass.edu/research/taems/white/.

[15] J. Huang and M. S. Fox. An Ontology of Trust – Formal Semantics and Transitivity. In *Proceedings of the Eighth International Conference on Electronic Commerce (ICEC'06)*, pages 259-270, 2006.

[16] A. Jøsang. A subjective metric of authentication. In *Proceedings of the 5th European Symposium on Research in Computer Security (ESORICS'98)*, Springer-Verlag, 1998.

[17] A. Jøsang, R. Hayward, and S. Pope. Trust Network Analysis with Subjective Logic. In *Proceedings of the Australasian Computer Science Conference (ACSC'06)*, 2006.

[18] A. Jøsang and R. Ismail. The Beta Reputation System. In *Proceedings of the 15th Bled Conference on Electronic Commerce*, pages 17-19, 2002.

[19] D. Khanafiah and H. Situngkir. Social Balance Theory: Revisiting Heider's Balance Theory for many agents. Bandung Fe Institute, Jawa Barat, Indonesia, 2004. http://arxiv.org/abs/nlin.PS/0405041.

[20] V. Lesser, K. Decker, T. Wagner, N. Carver, A. Garvey, B. Horling, D. Neiman, R. Podorozhny, M. N. Prasad, A. Raja, R. Vincent, P. Xuan, and X. Q. Zhang. Evolution of the GPGP/TÆMS Domain-Independent Coordination Framework. *Autonomous Agents and Multi-Agent Systems*, 9(1-2):87-143, 2004.

[21] H. Luo, J. Tao, and Y. Sun. Entropy-based Trust Management for Data Collection in Wireless Sensor Networks. In *Proceedings of the 5th International Conference on Wireless communications, networking and mobile computing (WiCOM'09)*, IEEE, 2009.

[22] H. V. D. Parunak. Real-Time Agent Characterization and Prediction. In *Proceedings of International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'07), Industrial Track*, pages 1421-1428, ACM, 2007.

[23] H. V. D. Parunak. Interpreting Digital Pheromones as Probability Fields. In *Proceedings of the 2009 Winter Simulation Conference*, pages 1059-1068, 2009.

[24] H. V. D. Parunak. Pheromones, Probabilities, and Multiple Futures. In *Proceedings of the 11th International Workshop on Multi-Agent Based Simulation*, 2010.

[25] H. V. D. Parunak, T. Belding, R. Bisson, S. Brueckner, E. Downs, R. Hilscher, and K. Decker. Stigmergic Modeling of Hierarchical Task Networks. In *Proceedings of the Tenth International Workshop on Multi-Agent-Based Simulation (MABS 2009, at AAMAS 2009)*, pages 98-109, Springer, 2009.

[26] H. V. D. Parunak and S. Brueckner. Concurrent Modeling of Alternative Worlds with Polyagents. In *Proceedings of the Seventh International Workshop on Multi-Agent-Based Simulation (MABS06, at AAMAS06)*, pages 128-141, Springer, 2006.

[27] J. S. Sichman and R. Conte. Multi-Agent Dependence by Dependence Graphs. In *Proceedings of AAMAS '02*, 2002.

[28] J. S. Sichman, Y. Demazeau, R. Conte, and C. Castelfranchi. A Social Reasoning Mechanism Based on Dependence Networks. In *Proceedings of 11th European Conference on Artificial Intelligence*, pages 416-420, John Wiley and Sons, 1994.

[29] M. P. Singh, B. Yu, and M. Venkatraman. Community-based service location. *Communications of the ACM*, 44(4):49-54, 2001.

[30] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636-13641, 2010.

[31] K. Thirunarayan, D. K. Althuru, C. A. Henson, and A. P. Sheth. A Local Qualitative Approach to Referral and Functional Trust. In *Proceedings of the The 4th Indian International Conference on Artificial Intelligence (IICAI-09), pp. , .* pages 574-588, 2009.

[32] Y. Wang and M. P. Singh. Evidence-Based Trust: A Mathematical Model Geared for Multiagent Systems. *ACM Transactions on Autonomous and Adaptive Systems*, 5(4), 2010.

[33] Y. Wang and M. P. Singh. Evidence-Based Trust: A Mathematical Model Geared for Multiagent Systems. *ACM Transactions on Autonomous and Adaptive Systems*, in press, 2011.

[34] P. Yolum and M. P. Singh. Service Graphs for Building Trust. 2004.

[35] B. Yu and M. P. Singh. Distributed reputation management for electronic commerce. *Computational Intelligence* 18(4):535–549, 2002.

Logical Definitions of Lying

Chiaki Sakama

Department of Computer and Communication Sciences
Wakayama University,
Sakaedani, Wakayama 640-8510, Japan
sakama@sys.wakayama-u.ac.jp

Abstract. This paper provides logical analyses of various definitions of *lying*. We first formulate twelve definitions of lying that appear in the literature of philosophy and were comprehensively studied by Mahon [14]. We use a propositional multi-modal logic that can represent belief and intention of agents. We then compare different definitions of lying and examine which one is best supported by both logically and empirically.

General Terms: theory. **Keywords**: lying, modal logic, reasoning.

1 Introduction

Understanding *what is lying* is necessary in identifying trustful agents, and providing ways to protect users from being deceived in multiagent societies. The problem has been studied in the field of philosophy and there is a number of different definitions of lying. Recently, James E. Mahon [14] provides a comprehensive study of lying, in which he examines twelve different definitions of lying that appear in the literature and argues which one is well-defined and empirically acceptable. He judges that one definition is acceptable or not by considering whether it satisfies four necessary conditions for lying, and by excluding those definitions that permit counter-intuitive examples. As a result, he concludes that two of the twelve definitions are the best ones.

There are several studies that provide formal account of lying. Bonatti et al. [1] study a theory of databases that could lie to users to preserve security. Firozabadi and Jones [9] define lying in terms of action logic. O'Neill [16] formulates various types of speech acts including lying using an epistemic logic. Sklar et al. [20] formulate lying with argument-based dialogues. Sakama et al. [17] provide logical analyses of different categories of dishonesties. These studies employ one definition of lying and provide its logical account for their purposes.

The goal of this paper is to provide logical formulation of different definitions of lying and compare their formal properties. We first formulate twelve definitions of lying that have been proposed in the philosophical literature and were analyzed in an informal way by Mahon [14]. To this end, we use a multi-modal logic that can represent belief and intention of agents. We then provide logical analyses for different definitions of lying and investigate their formal properties. We introduce five conditions that are requested to be satisfied by the definition of lying. The results of this paper provide formal grounds for the selection of best definitions of lying, and make us better understand what is lying.

2 A Logic for Belief and Intention

In this paper, we use a propositional modal logic of intentional communication [7]. A propositional modal language L_0 is built from a finite set of propositional constants $\{p, q, r, \ldots\}$ on the logical connectives $\neg$, $\vee$, $\wedge$, $\supset$, $\equiv$, and on two families of modal operators, $(B_a)_{a \in A}$ and $(I_a)_{a \in A}$, where A is a finite set of agents. Well-formed formulas (or *sentences*) in L_0 are defined as usual as those belonging to a multi-modal propositional logic. Sentences in L_0 will be denoted by the small Greek letters, and parentheses are employed as usual to clarify the structure of sentences. $\top$ and $\bot$ represent valid and contradictory sentences, respectively. The intuitive readings of $B_a\phi$ and $I_a\phi$ are that an agent a believes that ϕ and intends that ϕ, respectively. A Kripkean semantics is defined for L_0. Informally speaking, $B_a\phi$ (resp. $I_a\phi$) holds iff ϕ is true in all states of affairs compatible with a's current beliefs (resp. intentions). For example, if ϕ means that it rains, $I_a\phi$ should be read as "a intends to act in such a way that he or she brings about a state of affairs in which it rains".[1] A logic BI_0 is defined over L_0, that is an extension of KD45$_n$ [12] and has the following axioms and inference rules:

$(\mathbf{P})$ All propositional tautologies.
$(\mathbf{K_B})$ $B_a\phi \wedge B_a(\phi \supset \psi) \supset B_a\psi$ and $(\mathbf{K_I})$ $I_a\phi \wedge I_a(\phi \supset \psi) \supset I_a\psi$.
$(\mathbf{D_B})$ $B_a\phi \supset \neg B_a\neg\phi$ and $(\mathbf{D_I})$ $I_a\phi \supset \neg I_a\neg\phi$.
$(\mathbf{4_B})$ $B_a\phi \supset B_aB_a\phi$ and $(\mathbf{4_{IB}})$ $I_a\phi \supset B_aI_a\phi$.
$(\mathbf{5_B})$ $\neg B_a\phi \supset B_a\neg B_a\phi$ and $(\mathbf{5_{IB}})$ $\neg I_a\phi \supset B_a\neg I_a\phi$.

$$(\mathbf{MP})\ \frac{\phi \quad \phi \supset \psi}{\psi}, \qquad (\mathbf{N_B})\ \frac{\phi}{B_a\phi}, \qquad (\mathbf{N_I})\ \frac{\phi}{I_a\phi}.$$

Remark that $(\mathbf{N_I})$ says that all theorems hold at all state of affairs that an agent might intend to bring about [7].

To represent a speech act of an agent, we introduce the unary predicate $utter_{xy}$ defined over sentences in L_0 with $x, y \in A$. An expression $utter_{ab}(\sigma)$ means that an agent a expresses a sentence σ to an agent b. In particular, $utter_{aa}(\sigma)$ means that a expresses a sentence σ but the statement is directed to no one. A language L_0^U is defined as L_0 together with the predicate $utter_{xy}$. Sentences in L_0 are extended to L_0^U accordingly. If an agent utters something, he/she intends the speech act and is aware of his/her utterance. This is expressed by the next axiom:

$$(\mathbf{U_{IB}}) \qquad utter_{ab}(\sigma) \supset I_a(utter_{ab}(\sigma)) \wedge B_a(utter_{ab}(\sigma))\,.$$

We also assume that any utterance to a hearer is recognized by the hearer, and the speaker believes the recognition by the hearer. This is expressed by the axiom:

$$(\mathbf{U_{BB}}) \qquad utter_{ab}(\sigma) \supset B_b(utter_{ab}(\sigma)) \wedge B_aB_b(utter_{ab}(\sigma))\,.$$

The system BI_0^U, defined over L_0^U, is the weakest extension of BI_0 by the two axioms $(\mathbf{U_{IB}})$ and $(\mathbf{U_{BB}})$. If a sentence ϕ is a theorem of BI_0^U, we write $\vdash \phi$. Note that by $\mathbf{N_B}$ and $\mathbf{N_I}$, each agent believes and intends that other agents follow the same logic BI_0^U. Thus, $B_aB_b\phi \supset B_a\neg B_b\neg\phi$ and $B_a(I_b\phi \wedge I_b(\phi \supset \psi)) \supset B_aI_b\psi$, for instance.

[1] With such an interpretation, it becomes acceptable to apply an intention operator to arbitrary sentences [7].

3 Definitions of Lying

Mahon [14, 15] compares different definitions of lying in the philosophical literature and argues which definitions are most intuitive and acceptable. According to [15], there are at least four necessary conditions for lying. If a person lies, then **(i)** the person makes a statement (**statement condition**), and **(ii)** the person believes the statement to be false (**untruthfulness condition**), and **(iii)** the untruthful statement is made to another person (**addressee condition**), and **(iv)** the person intends that other person's believing the untruthful statement to be true (**intention to deceive addressee condition**). In what follows, we reformulate twelve definitions of lying that are argued in [14]. In this section, a and b represent two agents and σ and λ are two sentences of L_0^U.

3.1 Vrij's Definition

Vrij considers lying as "a successful or unsuccessful deliberate attempt, without forewarning, to create in another a belief which the communicator considers to be untrue" [21]. Vrij's definition is stated in [14] as follows:

> (L1) *To lie (to another person) is: to attempt to create a believed-false belief without forewarning (in another person).*

(L1) is formulated in L_0^U as follows.

Definition 3.1 (L1) $\quad L1_{ab}(\sigma) \overset{def}{=} B_a\neg\sigma \wedge I_aB_b\sigma \wedge B_aB_bB_a\sigma.$

(L1) represents that an agent a lies to another agent b on the sentence σ if a believes σ to be false and intends b's believing σ. Moreover, without forewarning, a believes that it is justified for b to believe that σ is believed to be true by a. (L1) does not satisfy the statement condition and the addressee condition. It satisfies the untruthfulness condition ($B_a\neg\sigma$) and the intention to deceive addressee condition ($I_aB_b\sigma$). Vrij's definition does not require any statement. Vrij says that "lying does not necessarily require the use of words. The athlete who fakes a foot injury after a bad performance is lying without using words" [21]. On the other hand, (L1) requires lying to be an act that happens without forewarning. Thus, "magicians are therefore not lying during their performance, as people in the audience expect to be deceived" (ibid). Mahon views (L1) too broad as a definition of lying. This is because "according to (L1), feigning a yawn, wearing a hairpiece, making a phony smile, wearing an engagement ring when one is not engaged, not wearing a wedding ring when one is married, or pretending to talk to someone on a cell phone, etc., is lying" [14]. Mahon concludes that (L1) is merely a definition of attempting to deceive and then rejects (L1) as a definition of lying.

3.2 Shibles's Definition

Shible's definition of lying [18] is stated in [14] as follows:

> (L2) *To lie (to another person) is: to make a believed-false statement (to another person).*

(L2) is formulated in L_0^U as follows.

Definition 3.2 (L2) $L2_{ab}(\sigma) \overset{def}{=} (utter_{aa}(\sigma) \vee utter_{ab}(\sigma)) \wedge B_a \neg\sigma.$

(L2) represents that an agent a lies on the sentence σ if a utters a believed-false sentence σ. (L2) satisfies the statement condition $(utter_{aa}(\sigma) \vee utter_{ab}(\sigma))$ but does not satisfy the addressee condition. (L2) also satisfies the untruthfulness condition $(B_a \neg\sigma)$ but does not satisfy the intention to deceive addressee condition. Different from (L1), (L2) requires that a statement be made (by uttering). This implies that, according to (L2), it is impossible for a person to lie by omitting to utter an expression [14]. On the other hand, (L2) does not require that the statement be made to anyone. Shible says in [18] that "a lie is merely a contradiction between belief (self-talk) and expression". Thus, "according to (L2), if a person goes into a room he believes to be empty, and utters untruthful statements, then that person is lying" [14]. Mahon rejects (L2) as "it does seem wrong that simply making an untruthful statement, to no one, is lying" (ibid).

3.3 Bok's Definition

Bok's definition of lying [2] is stated in [14] as follows:

> (L3) *To lie (to another person) is: to make a statement (to another person) with the intention to deceive (the other person).*

(L3) is formulated in L_0^U as follows.

Definition 3.3 (L3) $L3_{ab}(\sigma, \lambda) \overset{def}{=} utter_{ab}(\sigma) \wedge B_a B_b(\sigma \supset \lambda) \wedge B_a \neg\lambda \wedge I_a B_b \lambda.$

(L3) represents that an agent a lies to another agent b on the sentence σ if (i) a utters a sentence σ to another agent b, (ii) a believes that b's believing σ leads b to believing λ, (iii) a believes the falsity of λ, and (iv) believing λ by b is what a intends to achieve. (L3) satisfies both the statement condition and the addressee condition $(utter_{ab}(\sigma))$. (L3) also satisfies the intention to deceive addressee condition $(I_a B_b \lambda)$. On the other hand, (L3) does not satisfy the untruthfulness condition. That is, an agent a can utter a *believed-true* sentence σ with the intention that b uses it to reach a wrong conclusion λ. Mahon rejects (L3) "because it allows for the possibility of lying by making a truthful statement". According to this definition, if a father says to his child that "You will get a Christmas gift" and the statement makes his child believe that Santa Clause will come, then he is lying. In this case, however, it seems natural to consider that father does not lie because his child will get a Christmas gift, even if his intention is to make his child believe the existence of Santa Clause.

3.4 OED Definition

The *Oxford English Dictionary* defines a lie as "a false statement made with the intent to deceive". It is redefined in [14] as follows:

> (L4) *To lie (to another person) is: to make a false statement (to another person) with the intention to deceive (some person or other).*

(L4) is formulated in L_0^U as follows.

Definition 3.4 (L4) $L4_{ab}(\sigma, \lambda) \overset{def}{=} utter_{ab}(\sigma) \wedge \neg\sigma \wedge B_a B_b(\sigma \supset \lambda) \wedge B_a \neg\lambda \wedge I_a B_b \lambda.$

(L4) satisfies both the statement condition and the addressee condition ($utter_{ab}(\sigma)$). (L4) also satisfies the intention to deceive addressee condition ($I_a B_b \lambda$). On the other hand, (L4) does not satisfy the untruthfulness condition. It requires the statement to be false (**falsity condition**), rather than to be believed to be false. Mahon rejects (L4) because "one can lie by being truthful with an intention to deceive, when it just so happens that one is mistaken" [14]. For instance, if a clock strikes nine and a mother says to her child that "Nine! It's time to go to bed, otherwise, the devil will come". She expects that her child believes that the devil will come if he/she sits up late at night. If the clock is wrong and it is five to nine, then, according to (L4), she is lying. Else if the clock is correct, she is not lying. Thus, whether she lies or not depends on the correctness of the clock, and has no relation with devil's coming, which is unintuitive. There are other controversial examples and Mahon concludes that "it is better to say that one person is attempting to deceive another person, rather than to say that someone is lying to someone" (ibid).

3.5 Coleman and Kay's Definition

Coleman and Kay [6] strengthen the OED definition and require a lie to be a statement that is both false and believed-false. The definition is stated in [14] as follows:

> (L5) *To lie (to another person) is: to make a believed-false and false statement (to another person) with the intention that that statement be believed to be true (by the other person).*

(L5) is formulated in L_0^U as follows.

Definition 3.5 (L5) $L5_{ab}(\sigma) \overset{def}{=} utter_{ab}(\sigma) \wedge B_a \neg\sigma \wedge \neg\sigma \wedge I_a B_b \sigma.$

(L5) satisfies both the statement condition and the addressee condition ($utter_{ab}(\sigma)$). (L5) satisfies the untruthfulness condition ($B_a \neg\sigma$) as well as the falsity condition ($\neg\sigma$). (L5) also satisfies the intention to deceive addressee condition ($I_a B_b \sigma$). (L5) is considered a special case of (L4) with $\lambda \equiv \sigma$. Mahon rejects (L5) because "a person is not lying simply because the untruthful statement that she makes with the intention to deceive her addressee just happens to be true" [14], and "it does seem peculiar that whether or not one is lying depends upon luck" (ibid).

3.6 Kupfer's Definition

Kupfer [13] addresses that "a person lies when he asserts something to another which he believes to be false with the intention of getting the other to believe it to be true". It is redefined in [14] as follows:

> (L6) *To lie (to another person) is: to make a believed-false statement (to another person) with the intention that that statement be believed to be true (by the other person).*

(L6) is formulated in L_0^U as follows.

Definition 3.6 (L6) $L6_{ab}(\sigma) \overset{def}{=} utter_{ab}(\sigma) \wedge B_a\neg\sigma \wedge I_aB_b\sigma.$

(L6) satisfies both the statement condition and the addressee condition ($utter_{ab}(\sigma)$). (L6) also satisfies the untruthfulness condition ($B_a\neg\sigma$), and the intention to deceive addressee condition ($I_aB_b\sigma$). According to Mahon, "if any definition of lying may lay claim to being the standard definition of lying, then it is (L6)" [14]. (L6) represents an intention to deceive about the contents of the statement that is made. On the other hand, there would be an intention to deceive about one's belief in the truth of the statement that one makes (**believed truthfulness condition**) [15]. Borrowing an example of [14], suppose an FBI agent working undercover in a criminal organization. The crime boss notices this fact, but the FBI agent has no suspicion of this. If the crime boss says the FBI agent that there are no informants in his organization, then the boss cannot intend that the FBI agent believes this statement to be true because the boss knows that the agent is an informant. In this case, the crime boss can only intend that the FBI agent believes that the boss believes this statement to be true. According to (L6), the boss is not lying to the FBI agent. To cope with such cases, Mahon modifies (L6) as follows:

> (L6*) *To lie (to another person) is: to make a believed-false statement (to another person), either with the intention that that statement be believed to be true (by the other person), or with the intention that it be believed (by the other person) that that statement is believed to be true (by the person making the statement), or with both intentions.*

The modified version (L6*) is formulated in L_0^U as follows.

Definition 3.7 (L6*) $L6^*_{ab}(\sigma) \overset{def}{=} utter_{ab}(\sigma) \wedge B_a\neg\sigma \wedge (I_aB_b\sigma \vee I_aB_bB_a\sigma).$

(L6*) states that a lie requires either an intention to deceive about the contents of the statement that is made, or an intention to deceive about the beliefs of the person making the statement, or both. The believed truthfulness condition is represented by $I_aB_bB_a\sigma$. Mahon asserts that (L6) and (L6*) are two best definitions of lying [14].

3.7 Frankfurt's Definition

Frankfurt [10] also considers lying involves two distinct intentions to deceive: the one is "about the state of affairs to which he (a liar) explicitly refers and of which he is purporting to give a correct account", and the other is "about his own beliefs and what is going on in his mind" (ibid). Frankfurt's definition is stated in [14] as follows:

> (L7) *To lie (to another person) is: to make a believed-false statement (to another person) with the intention that that statement be believed to be true (by the other person) and with the intention that it be believed (by the other person) that that statement is believed to be true (by the person making the statement).*

(L7) is formulated in L_0^U as follows.

Definition 3.8 (L7) $L7_{ab}(\sigma) \overset{def}{=} utter_{ab}(\sigma) \wedge B_a\neg\sigma \wedge I_aB_b\sigma \wedge I_aB_bB_a\sigma.$

(L7) satisfies the statement condition, the addressee condition ($utter_{ab}(\sigma)$), and the untruthfulness condition ($B_a\neg\sigma$). (L7) satisfies both the intention to deceive addressee condition ($I_aB_b\sigma$) and the believed truthfulness condition ($I_aB_bB_a\sigma$). In contrast to (L6*), (L7) requires two distinct intentions to deceive be present and rules out cases in which only one intention to deceive is present. So, "if the lie works, then its victim is twice deceived" [10]. According to (L7), the crime boss is not lying to the FBI agent in the example given in Section 3.6. With this and another reasons, Mahon rejects (L7).

3.8 Chisholm and Feehan's Definition

Chisholm and Feehan [5] provide a complex definition of lying which is restated by [14] as follows.

> (L8) *To lie (to another person) is: to make a believed-not-true or believed-false statement (to another person), under conditions that are such that, (i) it is believed (by the person making the statement) that it is justified (for the other person) to believe that that statement is believed to be true (by the person making the statement), and (ii) it is believed (by the person making the statement) that it is justified (for the other person) to believe that it is intended (by the person making the statement) that it be believed (by the other person) that that statement is believed to be true (by the person making the statement).*

(L8) is formulated in L_0^U as follows.

Definition 3.9 (L8) $L8_{ab}(\sigma) \overset{def}{=} utter_{ab}(\sigma)\wedge(\neg B_a\sigma\vee B_a\neg\sigma)\wedge B_aB_bB_a\sigma\wedge B_aB_bI_aB_bB_a\sigma.$

(L8) satisfies both the statement condition and the addressee condition ($utter_{ab}(\sigma)$). However, (L8) does not always require the untruthfulness condition ($\neg B_a\sigma \vee B_a\neg\sigma$), and it does not require the intention to deceive addressee condition. Although the condition $B_aB_bB_a\sigma$ appears in (L1), the remaining conditions are quite different from those appearing in (L2)–(L7). $B_aB_bB_a\sigma \wedge B_aB_bI_aB_bB_a\sigma$ says that a believes that b believes that not only a's believing σ, but also a's intention to making b's believing $B_a\sigma$. According to (L8), an untruthful statement that is made merely in play or in irony is not a lie because the speaker does not believe that the hearer believes that the speaker believes what he says. On the other hand, (L8) has "the very odd and unacceptable result that a notoriously dishonest person cannot lie to people who he knows distrust him. Their definition implies that it is self-contradictory to say that I lie when I know that others know that I am lying" [4]. Mahon also rejects (L8) by similar reasons.

3.9 Simpson's Definition

Simpson [19] provides yet another definition of lying. It is stated in [14] as follows:

> (L9) *To lie (to another person) is: to make a believed-false statement (to another person) with the intention that that statement be believed to be true (by the other person), and with the intention that it be believed (by the other person) that that statement is believed to be true (by the person making the statement), and with the intention that it be believed (by the other person) that it is*

> *intended (by the person making the statement) that it be believed (by the other person) that that statement is believed to be true (by the person making the statement).*

(L9) is formulated in L_0^U as follows.

Definition 3.10 (L9) $L9_{ab}(\sigma) \overset{def}{=} utter_{ab}(\sigma) \wedge B_a \neg\sigma \wedge I_a B_b \sigma \wedge I_a B_b B_a \sigma \wedge I_a B_b I_a B_b B_a \sigma.$

(L9) satisfies the statement condition, the addressee condition ($utter_{ab}(\sigma)$), and the untruthfulness condition ($B_a \neg\sigma$). (L9) requires three different intentions: a intends to (i) make b believe the untruthful statement σ ($I_a B_b \sigma$), (ii) make b believe that a believes σ ($I_a B_b B_a \sigma$), and (iii) make b believe that a intends (ii) ($I_a B_b I_a B_b B_a \sigma$). Like (L7), (L9) requires the believed truthfulness condition. Hence, Mahon rejects (L9) by the same reason as (L7).

3.10 Carson's Definition

Carson [3] defines lying in terms of the warranty of the statement. It is defined in [14] as follows:

> (L10) *To lie (to another person) is: to make a not-believed-true, and false, statement (to another person), in a context in which the truth of the statement is thereby warranted (by the person making the statement) (to the other person), (the person making the statement) not believing that the truth of the statement is not being warranted (by the person making the statement) (to the other person).*

(L10) is formulated in L_0^U as follows.

Definition 3.11 (L10) $L10_{ab}(\sigma) \overset{def}{=} utter_{ab}(\sigma) \wedge \neg B_a \sigma \wedge \neg\sigma \wedge B_b B_a \sigma \wedge \neg B_a \neg B_b B_a \sigma.$

(L10) satisfies the statement condition and the addressee condition ($utter_{ab}(\sigma)$). On the other hand, (L10) does not require untruthfulness but requires falsity of the statement. Also, (L10) does not require an intention to deceive the addressee. The condition $B_b B_a \sigma$ represents the situation that a speaker warrants the truth of his/her statement to the hearer. According to (L10), a can lie only if b believes that a believes σ. Mahon argues that "if one makes an untruthful and false statement to an audience, and if one intends to warrant the truth of that statement to one's audience, then one is lying, even if, unbeknownst to one, the context is such that one's audience does not take one to be warranting the truth of one's statement to them, and hence, one's audience does not believe one's statement to be true" [14]. Mahon then rejects (L10).

3.11 Fallis's Definition

We finally provide a definition by Fallis [8], which differs from most other definitions of lying given so far. Mahon provides it as follows:

> (L11) *To lie (to another person) is: to make a believed-false statement (to another person) while believing that the context is one in which the norm 'Do not say what you believe to be false' is in effect.*

In the above, ‘Do not say what you believe to be false’ is called the *Gricean conversational norm of truthfulness* [11]. (L11) is formulated in L_0^U as follows.

Definition 3.12 (L11) $L11_{ab}(\sigma) \overset{def}{=} utter_{ab}(\sigma) \wedge B_a\neg\sigma \wedge B_aB_b(utter_{ab}(\sigma) \supset \neg B_a\neg\sigma)$.

(L11) satisfies the statement condition and the addressee condition ($utter_{ab}(\sigma)$). (L11) also satisfies the untruthfulness condition ($B_a\neg\sigma$), but does not satisfy the intention to deceive addressee condition. In (L11), the norm of truthfulness is represented by $B_aB_b(utter_{ab}(\sigma) \supset \neg B_a\neg\sigma)$. Thus, “unlike (L10), whether or not the context is one in which the truthfulness norm ‘Do not say what you believe to be false’ is in effect is determined entirely by the beliefs of the person making the statement” [14]. Thus, if one makes an untruthful statement to an audience while believing that the truthfulness norm is in effect, then one is lying even if one’s audience does not believe one’s statement to be true. With this respect, Mahon considers that (L11) is better than (L10). On the other hand, Mahon argues that (L11) has a problem when one “make(s) an untruthful statement, while believing that one is in a context in which the norm of truthfulness is in effect, and not intend that one’s untruthful statement be believed to be true” [14]. For instance, suppose that a policeman asks a drunken man “Where is your home?” and the drunk replies “On the moon”. The policeman then considers that the drunk is just kidding, rather than lying. With this reason, Mahon does not consider (L11) appropriate.

4 Comparison between Different Definitions

As addressed in Section 3, Mahon considers four necessary conditions for lying: statement condition, untruthfulness condition, addressee condition, and intention to deceive addressee condition, as well as other (not necessarily required) conditions such as falsity condition (L4, etc) and believed truthfulness condition (L6*, etc). Here we introduce five additional conditions that are considered natural to be satisfied by the definition of lying.[2]

- Lying on valid sentences is impossible (**inability to lie on valid sentences**, or **inability**-$\top$ for short).
- Lying on contradictory sentences is impossible (**inability to lie on contradictory sentences**, or **inability**-$\bot$ for short).
- Lying on two sentences σ and $\neg\sigma$ at the same time is impossible (**inability to lie on mutually complementary sentences**, or **inability**-$\neg$ for short).
- A liar is aware of his/her dishonest act (**awareness**).
- Lying to oneself leads to contradiction (**self-contradiction**).

We examine whether twelve definitions of lying satisfy the above five conditions. In what follows, $LIE_{ab}(\sigma)$ means one of the definitions (L1), (L2), and (L5)–(L11). We prove propositions for $LIE_{ab}(\sigma)$, but the same proofs are applied for (L3) and (L4) that have the additional parameter λ.

[2] The four conditions, inability-$\top$, inability-$\bot$, awareness, and self-contradiction, are also considered in [17].

Proposition 4.1 (inability-$\top$) $\vdash LIE_{ab}(\top) \supset \bot$ *holds if* $LIE_{ab}(\sigma)$ *includes either* $B_a\neg\sigma$ *or* $\neg B_a\sigma$ *or* $\neg\sigma$.

Proof. If $LIE_{ab}(\sigma)$ includes either $B_a\neg\sigma$ or $\neg B_a\sigma$, $LIE_{ab}(\top)$ implies $B_a\bot$ that implies $\neg B_a\top$ ($\mathbf{D_B}$), while $\top$ implies $B_a\top$ ($\mathbf{N_B}$). Contradiction. If $LIE_{ab}(\sigma)$ includes $\neg\sigma$, $LIE_{ab}(\top)$ implies $\bot$. □

By Proposition 4.1, we can see that (L1), (L2), (L4), (L5), (L6), (L6*), (L7), (L8), (L9), (L10), and (L11) satisfy the property of inability to lie on valid sentences. By contrast, $L3_{ab}(\top, \lambda)$ does not imply $\bot$.

Proposition 4.2 (inability-$\bot$) $\vdash LIE_{ab}(\bot) \supset \bot$ *holds if* $LIE_{ab}(\sigma)$ *includes either* $I_aB_b\sigma$ *or* $I_aB_bB_a\sigma$ *or* $B_aB_bB_a\sigma$.

Proof. If $LIE_{ab}(\sigma)$ includes $I_aB_b\sigma$, then $LIE_{ab}(\bot)$ implies $I_aB_b\bot$, while $B_b\top$ implies $\neg B_b\bot$ ($\mathbf{D_B}$) that implies $I_a\neg B_b\bot$ ($\mathbf{N_I}$) then $\neg I_aB_b\bot$ ($\mathbf{D_I}$). Contradiction.

If $LIE_{ab}(\sigma)$ includes $I_aB_bB_a\sigma$, then $LIE_{ab}(\bot)$ implies $I_aB_bB_a\bot$. On the other hand, $B_a\top$ implies $\neg B_a\bot$ ($\mathbf{D_B}$) that implies $B_b\neg B_a\bot$ ($\mathbf{N_B}$) then $\neg B_bB_a\bot$ ($\mathbf{D_B}$). Thus, $I_a\neg B_bB_a\bot$ ($\mathbf{N_I}$), so $\neg I_aB_bB_a\bot$ ($\mathbf{D_I}$) which contradicts $I_aB_bB_a\bot$.

If $LIE_{ab}(\sigma)$ includes $B_aB_bB_a\sigma$, then $LIE_{ab}(\bot)$ implies $B_aB_bB_a\bot$. On the other hand, $B_a\top$ implies $\neg B_bB_a\bot$ as above, which implies $B_b\neg B_bB_a\bot$ ($\mathbf{N_B}$) then $\neg B_bB_bB_a\bot$ ($\mathbf{D_B}$). Contradiction. □

By Proposition 4.2, we can see that (L1), (L5), (L6), (L6*), (L7), (L8), (L9), and (L10) satisfy the property of inability to lie on contradictory sentences. On the other hand, $L11_{ab}(\bot)$ implies $B_aB_b(utter_{ab}(\bot) \supset \neg B_a\top)$, hence $B_a(B_b(utter_{ab}(\bot)) \supset B_b\neg B_a\top)$ and $B_aB_b(utter_{ab}(\bot)) \supset B_aB_b\neg B_a\top$ ($\mathbf{K_B}$). Since $utter_{ab}(\bot)$ implies $B_aB_b(utter_{ab}(\bot))$ ($\mathbf{U_{BB}}$), it holds that $B_aB_b\neg B_a\top$ by (**MP**). As $B_a(B_b\neg B_a\top \supset \neg B_bB_a\top)$, it holds that $B_a\neg B_bB_a\top$ thereby $\neg B_aB_bB_a\top$ ($\mathbf{D_B}$). This contradicts $B_aB_bB_a\top$ that is obtained from $\top$ by iteratively applying ($\mathbf{N_B}$). Hence, $L11_{ab}(\bot)$ also imply $\bot$. By contrast, $L2_{ab}(\bot)$, $L3_{ab}(\bot, \lambda)$, and $L4_{ab}(\bot, \lambda)$ do not imply $\bot$.

Proposition 4.3 (inability-$\neg$) $\vdash LIE_{ab}(\sigma) \wedge LIE_{ab}(\neg\sigma) \supset \bot$ *holds if* $LIE_{ab}(\sigma)$ *includes either* $B_a\neg\sigma$ *or* $\neg\sigma$.

Proof. If $LIE_{ab}(\sigma)$ includes $B_a\neg\sigma$, $LIE_{ab}(\sigma) \wedge LIE_{ab}(\neg\sigma)$ implies $B_a\neg\sigma \wedge B_a\sigma$. $B_a\neg\sigma$ implies $\neg B_a\sigma$ ($\mathbf{D_B}$). Contradiction. If $LIE_{ab}(\sigma)$ includes $\neg\sigma$, $LIE_{ab}(\sigma) \wedge LIE_{ab}(\neg\sigma)$ implies $\neg\sigma \wedge \sigma$. Contradiction. □

By Proposition 4.3, we can see that (L1), (L2), (L4), (L5), (L6), (L6*), (L7), (L9), (L10) and (L11) satisfy the property of inability to lie on mutually complementary sentences. On the other hand, (L3) and (L8) do not satisfy the property.

Proposition 4.4 (awareness) $\vdash LIE_{ab}(\sigma) \supset B_a(LIE_{ab}(\sigma))$ *holds for any sentence* σ *for the definitions of (L1), (L2), (L3), (L6), (L6*), (L7), (L8), (L9), and (L11).*

Proof. The result holds by each definition and the axioms ($\mathbf{U_{IB}}$), ($\mathbf{4_B}$), and ($\mathbf{4_{IB}}$). □

(L4), (L5) and (L10) do not satisfy the property because they contain the condition of falsity, that is, $\neg\sigma$ does not imply $B_a\neg\sigma$.

Table 1. Comparison of twelve definitions

	L1	L2	L3	L4	L5	L6	L6*	L7	L8	L9	L10	L11
statement*		√	√	√	√	√	√	√	√	√	√	√
addressee*		(√)	√	√	√	√	√	√	√	√	√	√
untruthful*	√	√			√	√	√	√	(√)	√		√
intention*	√		√	√	√	√	√	√		√		
believed truthful							(√)	√		√		
falsity				√	√						√	
inability-$\top^\dagger$	√	√		√	√	√	√	√	√	√	√	√
inability-$\bot^\dagger$	√				√	√	√	√	√	√	√	√
inability-$\neg^\dagger$	√	√		√	√	√	√	√		√	√	√
awareness†	√	√	√			√	√	√	√	√		√
self-contradiction†	√		√	√	√	√	√	√	√	√	√	√

$*$: necessary conditions by Mahon; †: conditions considered in this paper.

Proposition 4.5 (self-contradiction) $\vdash LIE_{aa}(\sigma) \supset \bot$ *holds for any sentence* σ *if* $LIE_{ab}(\sigma)$ *includes either* $B_a\neg\sigma \wedge I_aB_b\sigma$ *or* $B_a\neg\sigma \wedge I_aB_aB_a\sigma$ *or* $\neg B_a\sigma \wedge B_bB_a\sigma$.

Proof. If $LIE_{ab}(\sigma)$ includes $B_a\neg\sigma \wedge I_aB_b\sigma$, then $LIE_{aa}(\sigma)$ implies $B_a\neg\sigma \wedge I_aB_a\sigma$. $B_a\neg\sigma$ implies $\neg B_a\sigma$ ($\mathbf{D_B}$), which implies $I_a\neg B_a\sigma$ ($\mathbf{N_I}$). On the other hand, $I_aB_a\sigma$ implies $\neg I_a\neg B_a\sigma$ ($\mathbf{D_I}$). Contradiction.

If $LIE_{ab}(\sigma)$ includes $B_a\neg\sigma\wedge I_aB_bB_a\sigma$, then $LIE_{aa}(\sigma)$ implies $B_a\neg\sigma\wedge I_aB_aB_a\sigma$. $B_a\neg\sigma$ implies $\neg B_a\sigma$ ($\mathbf{D_B}$) that implies $B_a\neg B_a\sigma$ ($\mathbf{5_B}$) then $\neg B_aB_a\sigma$ ($\mathbf{D_B}$). Thus, $I_a\neg B_aB_a\sigma$ ($\mathbf{N_I}$), so $\neg I_aB_aB_a\sigma$ ($\mathbf{D_I}$) which contradicts $I_aB_aB_a\sigma$.

If $LIE_{ab}(\sigma)$ includes $\neg B_a\sigma \wedge B_bB_a\sigma$, then $LIE_{aa}(\sigma)$ implies $\neg B_a\sigma \wedge B_aB_a\sigma$. $\neg B_a\sigma$ implies $B_a\neg B_a\sigma$ ($\mathbf{5_B}$), which implies $\neg B_aB_a\sigma$ ($\mathbf{D_B}$). Contradiction. □

By Proposition 4.5, we can see that (L1), (L3), (L4), (L5), (L6), (L6*), (L7), (L9), and (L10) satisfy the property of self-contradiction. It is easy to see that (L8) and (L11) also satisfy the property, but (L2) is not.

Table 1 compares twelve definitions of lying from the viewpoint of satisfaction of various conditions explained so far. In the table, √ means satisfaction of each condition, so (L1), for instance, satisfies conditions of untruthfulness, intention to deceive addressee, inability-$\top$, inability-$\bot$, inability-$\neg$, awareness, and self-contradiction. (√) means that the condition is included as a disjunct.

By the table, we can observe that (L7) and (L9) satisfy most conditions. (L5), (L6), and (L6*) follow them. (L5) does not satisfy the awareness condition, which appears unintuitive. Mahon argues that believed truthfulness is often too strong. As a result, Mahon concludes that (L6) and (L6*) are best definitions among the twelve definitions. Since (L6) and (L6*) also satisfy the additional five conditions, inability-$\top$, inability-$\bot$, inability-$\neg$, awareness, and self-contradiction, we also support these two definitions.

Figure 1 shows relationship between different definitions of lying. In the figure, $X \to Y$ represents that X implies Y under the logic L_0^U. In the figure, we can observe that (L4) and (L9) are relatively stronger, while (L2) is relatively weaker. On the other hand, (L1), (L8) or (L10) has no implication relations with other definitions.

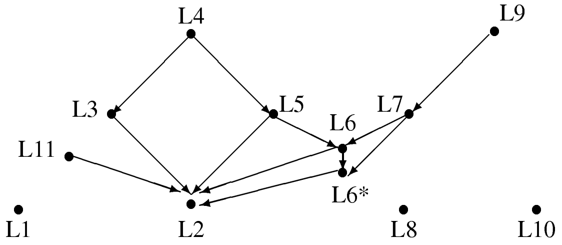


Fig. 1. Relationship between twelve definitions

5 Conclusion

In this paper, we gave logical definitions of lying and analyzed their formal properties. The results of this paper provide a formal ground for Mahon's informal argument that Kuper's definition (L6) and its modification (L6*) are most intuitive. The logic of belief and intention used in this paper is simple but expressive for abstracting the act of lying. The five new conditions that logically justify the act of lying, together with four necessary conditions by Mahon that empirically support lying, serve as criteria for judging whether yet another definition of lying is appropriate or not.

References

1. Bonatti, P. A., Kraus, S. and Subrahmanian, V. S.: Foundations of secure deductive databases. *IEEE Transactions on Knowledge and Data Engineering* 7(3), 406–422 (1995)
2. Bok, S.: *Lying: Moral Choice in Public and Private Life.* Random House (1978)
3. Carson, T. L.: The definition of lying. *Noûs* 40(2), 284–306 (2006)
4. Carson, T. L.: *Lying and deception: theory and practice.* Oxford University Press (2010)
5. Chisholm, R. M. and Feehan, T. D.: The intent to deceive. *Journal of Philosophy* 74(3), 143–159 (1977)
6. Coleman, L. and Kay, P.: Prototype semantics: the English word lie. *Language* 57(1), 26–44 (1981)
7. Colombetti, M.: A modal logic of intentional communication. *Mathematical Social Sciences* 38, 171–196 (1999)
8. Fallis, D.: What is lying? *Journal of Philosophy* 106(1), 29–56 (2009)
9. Firozabadi, B. S. and Jones, A. J. I.: On the characterisation of a trusting agent – aspects of a formal approach. In: C. Castelfranchi and Y. H. Tan (eds.), *Trust and Deception in Virtual Societies*, Kluwer Academic Publishers, pp. 157–168 (2001)
10. Frankfurt, H. G.: The faintest passion. In: *Necessity, Volition and Love.* Cambridge University Press (1992)
11. Grice, H. P.: *Studies in the Ways of Words.* Harvard University Press (1989)
12. Halpern, J. and Moses, J.: A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence* 54, 349–379 (1992)
13. Kupfer, J.: The moral presumption against lying. *Review of Metaphysics* 36, 103–126 (1982)
14. Mahon, J. E.: Two definitions of lying. *J. Applied Philosophy* 22(2), 211–230 (2008)
15. Mahon, J. E.: The definition of lying and deception. *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/lying-definition/ (2008)

16. O'Neill, B.: A formal system for understanding lies and deceit. *Jerusalem Conference on Biblical Economics* (2003)

17. Sakama, C., Caminada, M. and Herzig, A.: A logical account of lying. In: *Proc. 12th European Conference on Logics in Artificial Intelligence, Lecture Notes in Artificial Intelligence* 6341, pp. 286–299 (2010)

18. Shibles, W.: *Lying: A Critical Analysis.* The Language Press (1985)

19. Simpson, D.: Lying, liars and language. *Philosophy and Phenomenological Research* 52: 623–629 (2007)

20. Sklar, E., Parsons, S. and Davies, M.: When is it okay to lie? A simple model of contradiction in agent-based dialogues. *Argumentation in Multi-Agent Systems, Lecture Notes in Computer Sciences* 3366, pp. 251–261, Springer (2005)

21. Vrij, A.: *Detecting lies and deceit: the psychology of lying and the implications for professional practice.* John Wiley & Sons (2000)

Reasoning with Categories for Trusting Strangers: a Cognitive Architecture

Matteo Venanzi[1,2], Michele Piunti[1], Rino Falcone[1] and Cristiano Castelfranchi[1]
mv1g10@ecs.soton.ac.uk
{michele.piunti,rino.falcone,cristiano.castelfranchi}@istc.cnr.it

[1] Goal T^3 Group
Institute of Cognitive Science and Technologies (ISTC-CNR) Roma, Italy

[2] IAM Group
Univeristy of Southampton, Southampton, SO17 1BJ, UK

Abstract. A crucial issue for agents in open systems is the ability to filter out information sources in order to build an image of their counterparts, upon which a subjective evaluation of trust as a promoter of interactions can be assessed. While typical solutions discern relevant information sources by relying on previous experiences or reputational images, this work presents an alternative approach based on the cognitive ability to: (*i*) analyze heterogeneous information sources along different dimensions; (*ii*) ascribe qualities to unknown counterparts based on reasoning over abstract classes or categories; and, (*iii*) learn a series of emergent relationships between particular properties observable on other agents and their effective abilities to fulfill tasks. A computational architecture is presented allowing cognitive agents to dynamically assess trust based on a limited set of observable properties, namely explicitly readable signals (*Manifesta*) through which it is possible to infer hidden properties and capabilities (*Krypta*), which finally regulate agents' behavior in concrete work environments. Experimental evaluation discusses the effectiveness of trustor agents adopting different strategies to delegate tasks based on categorization.

1 Introduction

Interaction and *openness* are topics deserving the attention of the research agenda in Multi Agent Systems (MAS): interaction being at the basis of communication, coordination and cooperation, like for instance in virtual societies and networks; openness being at the basis of many of the applicative domains currently developed, like for instance open marketplaces characterized by an ecosystem of mobile devices, services and thousands of exploitable titles and applications. As indicated by many approaches, trust is a pivotal aspect for both interaction and openness. Trust is fundamental for facing the uncertainties typical of open societies, where heterogenous entities are forced to choose whether to interact or not with possibly unknown counterparts. Besides, being at the basis of any interplay, trust is a glue for the whole society: it can promote or prevent interactions of multiple entities, possibly governed by autonomous objectives and capabilities. Even more, trust plays a central role in decision making: it is diriment factor in deciding whether to externalize or not a given activity, or in deciding if a given task can be profitably delegated to another agent.

The downside of trust is that managing it is a costly process for agents. There is a problem of *trust formation*: in order to exploit the benefits of trust, agents need to build a knowledge model able to assess the trustworthiness for each possible counterpart, thus processing additional information about the others. A main issue is in filtering the information sources and in providing a mechanism for evaluating trust on such a basis. Existing literature suggests a couple of alternatives to an agent for assessing trust [7]. The first approach assumes to exploit *personal experience* to analyze how a given agent has performed in past interactions. Otherwise, the shared opinion circulating about a given agent could be exploited in terms of *recommendations/reputation*. In this paper we explore an alternative approach, based on the *reasoning/inference* about the others based on categories of agents. In this direction, we propose categorial trust as a suitable approach to trust formation, and we propose a series of computational mechanisms realizing it in cognitive agents.

Based on a socio-cognitive model of trust [5], we assume that for rationally trusting someone we need a theory of its mind (in case of a cognitive agent) or of its functioning (in case of a more simple artifact). Categorial trust is inspired to an heuristics commonly exploited by humans. It considers the cognitive ability to represent group behavior using general classes or categories of individuals, where categories can be shaped on a specific set of observable features and qualities. The claim of this work is to show that, as in the human case, considering an unknown agent as belonging to a known category allows to infer (or at least attribute) specific internal features for such unknown agent, not directly observable. This means to identify a set of agent's internal features determining how that agent will perform in specific situations. On such a basis, agents may recognize the strict correlation between the internal features of a possible trustee and its pragmatic performances in concrete tasks. In this sense the model recalls the notions of *Krypta* and *Manifesta* [1], according to which manifesta are observable signs for agents' krypta, a sort of internal properties ("qualities", "virtues" or "powers") exploitable to predict/explain their behaviors on specific tasks or activities. Categorial reasoning is provided in order to implement two different level of inference: the former, based on the agentive-personal level, allowing to refine the real capabilities of a given agent based on the analysis of its observable attributes; the latter, based on the societal-categorial level, allowing to refine or create new categories based on the appraised relation between the ability to fulfill a given task and the observable properties belonging to that class of agents. The model proposed in this paper will enable agents to work in both the levels of inference, being part of a cognitive architecture enabling agents to: (i) ascribe the effectiveness of a given category for a given task, thus identifying the right trustee on the basis of his potential categorization as expressed by its observable manifesta; (*ii*) assess trust towards a population of unknown agents in dynamic environment conditions, with tasks characterized by changing requirements; (*iii*)assess trust based on partial information about heterogeneous population of agents: a trustor only knows few manifesta for a given trustee.

The rest of the paper is organized as follows. Section 2 surveys related works focusing on the socio-cognitive approach to trust. Section 3 places the research challenge in terms of categorial trust, while Section 4 formalizes a cognitive architecture realizing it and describes a concrete programming model for its implementation. Section 5 presents simulative experiments and results aimed at evaluating the effectiveness of different trust formation strategies. Finally, Section 6 provides final discussion and perspectives.

2 Trusting Agents in Open Systems

Establishing trust in open system requires to effectively build a behavioral model of entities which typically are not known in advance (strangers). From an agent perspective, assessing trust is related to the problem of trust formation, which in open systems refers to the problem to filter a wide spectrum of information distributed within heterogenous sources. Several approaches to trust have been explored in MAS based on experience and reputation [7]. A first strategy relies on the ability to store information of past experiences, and build on such a personal knowledge a subjective model of trust. The same idea has been exploited to assess trust based on statistical analysis [12]. The weakness of these approaches is related the costs in terms of resources needed to explore the whole set of available options before having a direct experience on each available agent. Reputational approaches make use of shared information sources, like certified authorities, reputation and reports. Among others, Sabater at al. proposed a model based on agents' images and reputation [13], according to which social evaluations circulate and are represented as reported evaluations, which are exploited to promote trust formation. Other approaches, as the one explored for instance by [9], makes use of infrastructures making available certified reputation related to each possible trustee agent.

The suggestion to exploit categorial knowledge to assess trust is not new, and it has been theoretically explored for ascertain beforehand the trustworthiness of possible unknown counterparts [2]. In the context of computational models, the work by Wojcik et al. introduced the notion of prejudice filters to perceive particular trustees attributes [14]. Rules are extracted to avoid distrusted interactions, thus denying transactions which may be expected as not profitable. The Stereotrust approach proposed by Brunett et al. allow agents to build stereotypes based on the analysis of past interaction outcomes [4]. Data mining techniques are used to dynamically create classifiers based on personal knowledge. Classifiers are then applied to establish trustworthiness of possible trustees in absence of personal information. As explained in the next sections, the model proposed in this paper revises and extends the use of prejudices and stereotypes in the context of a more general theory of cognitive trust.

The socio-cognitive approach proposed by Castelfranchi and Falcone [5] considers trust as a cognitive process characterized by both relational and graded notions. A pivotal aspect of the socio-cognitive model is that trust formation is a cognitive process based on a series of cognitive ingredients through which the trustor evaluates the trustee in a specific environmental context, by assessing a particular configuration of (positive) expectation and reliance. Trust is a relational notion between a trustor agent (trust giver, ag_i) and a trustee agent (trust receiver, ag_j) which can be established in a given context C, and, most important, about a defined activity or task to be fulfilled (τ):

$$Trust(ag_i, ag_j, C, \tau)$$

. Accordingly, trust is a *graded* construct, and the degree of trust (DoT) comes from the degree of a series of cognitive ingredients, which can be resumed in terms of trustor's beliefs and goals. Summing up, an agent ag_i trusts ag_j about the task τ if DoT overcomes a given threshold σ:

$$DoT_{ag_i, ag_j, \tau} > \sigma$$

Within a group of possible trustees, we assume the trustor will prefer the one having the higher DoT. We omit for simplicity the characterization of trust in terms of additional facts that ag_i has to believe about the trustee and the external conditions (the interested reader can find formalized the approach in concrete implementations, as in [8]). In the particular approach described in this work, such a trustor's beliefs can be assumed as already established once the trustor is able to fill a given trustee in a given category (or class) of agents. Analyzing the wide spectrum of information sources allows ag_i to assess of a series of expectations on ag_j, which in turn makes it possible to assess trust and anticipate its behavior. In this view, trust formation can be assessed on the particular ability of ag_i to analyze a series of ag_j's observable properties (*Manifesta*) and, on such a basis, to infer a theory of ag_j mind (*Krypta*).

3 Cognitive Trust Formation

The approach to cognitive trust proposed in this work assumes two different level of reasoning: the *personal level* which allows to use the information available on the individual trustees, and the *categorial level*, related to the relationship between agents and their categories. Accordingly, for each possible trustee in the system we assume three types of observable information (manifesta). *Professional* and *dispositional* manifesta summarizes internal factors of trust attribution, related in particular to abilities and willingness of a given agent. These features can be exploited at a personal level, i.e., for ascribing a given agent in a specified (professional or dispositional) category. As humans normally do, a particular apparel, particular attitudes or situations can be exploited to find people playing a given role (i.e. a doctor, a dentist, a surgeon) or having a given attitude (i.e. careful, cautious, impulsive). The third class of manifesta considers the information not directly related to professional abilities and willingness, for example being male or female, old or young, religious or atheist, etc. We define this class as "crosscutting" manifesta. In the case of crosscutting manifesta, the relationship with agents krypta has to be *learned* at a categorial level. This is why, for instance, humans form the prejudice that being young, or female, or religious is a better category for fulfilling a series of activities. Summing up, each trustee present in the agent system is assumed as a carrier of three observable properties observable manifesta. For instance a trustee may present features as $\langle$*Surgeon, Cautious, Male* $\rangle$ or $\langle$*Pediatrician, Careful, Female* $\rangle$.

On such a structures, the objective to assess trust is twofold: on the one side it aims to give agents the ability to reason either on the personal level (direct experience), and on the categorial level (categorial experience); on the other side, it aims to show a model of trust built on various levels of information: personal and categorial. We envisage that such an approach may provide an effective heuristic to agents acting in open societies, where the information of prior direct transactions are scarce, and where the possibility to build trust models based on direct experience is infeasible.

In order to design a cognitive model general enough to develop different trust formation strategies, an open scenario has been envisaged. Autonomous agents have to cooperate to carry out a series of tasks inspired to a medical domain, and we assume agents playing two possible roles: patients and medical doctors. At each round, we assume that the *tasks*, inspired by medical diseases, are delegated by patients to doctors. We further assume doctor agents as allowed to enter and exit the system at each time step, thus characterizing the application domain as an *open* system.

3.1 Tasks

The set $\mathcal{T}$ indicates a set of tasks to be fulfilled by patients: $\mathcal{T} = \{\tau_1, \tau_2, ...\tau_N\}$. Each task is characterized by a list of *requirements* needed for its fulfillment: $\tau_j = \langle \tau_{id}, \tau_{Prof}, \tau_{Disp}, \tau_{Cross}, \tau_{State} \rangle$, where $1 \leq j \leq N$ and where requirements are shaped on various dimensions:

- $\tau_{Prof} = \{\alpha_{spec}, \alpha_1, ...\alpha_O\}$ defines abilities (professional) needed to fulfill the task. We assume in particular $\alpha_{spec} \in \tau_{Prof}$ as the pivotal requirement characterizing the task;
- $\tau_{Disp} = \{\omega_1, \omega_2, ...\omega_P\}$ defines willingness (dispositional) to fulfill the task;
- $\tau_{Cross} = \{\kappa_1, \kappa_2, ...\kappa_Q\}$ defines requirements that are not uniquely and immediately related to abilities and dispositions (crosscutting);

Table 1 (a) shows `Chickenpox` and `Appendicitis` as concrete examples of task specification.Task representation includes the structures related to dispositional, professional, and crosscutting categorial requirements. In the `Chickenpox` example, we assume that a specific requirement, called α_{spec}, is the pivotal one to fulfill the task. For instance, to fulfill the `Chickenpox` task, an α_{spec} *pediatr_spec* is needed in order to achieve a result value greater than 0.5. Notice that we assume the cross categorial attribute of being "female" as a task requirement. This

Chickenpox	
Abilities	
pediatr_spec	99
manual	90
literature	80
technique	90
Dispositions	
availability	90
caution	80
attention	70
Cross	
female	*true*

Appendicitis	
Abilities	
surgery_spec	99
manual	90
literature	50
technique	90
Dispositions	
availability	90
caution	90
attention	60
Cross	
male	*true*

Male
Crosscutting

Female
Crosscutting

Pediatrician	
Professional	
pediatr_spec:	[99 ... 100]
manual:	[70 ... 100]
literature:	[60 ... 100]
technique:	[70 ... 100]

Surgeon	
Professional	
surgery_spec:	[99 ... 100]
manual:	[75 ... 100]
literature:	[60 ... 100]
technique:	[60 ... 100]

Available	
Dispositional	
caution:	[50 ... 70]
attention:	[50 ... 70]
availability:	[60 ... 80]

Careful	
Dispositional	
caution:	[80 ... 100]
attention:	[90 ... 100]
availability:	[40 ... 60]

a) Tasks b) Crosscutting cat. c) Professional cat. d) Dispositional cat.

Table 1. Examples of Tasks and Categories specified in a medical domain.

means that, once the task can be fulfilled with a graded result, the contribute of being female consist in an improved outcome, once the fulfillment of a given task ranges from 0 to 100. In concrete implementation, each requirement is modeled as a threshold to be reached by an agent capability in order to be fulfilled[3].

3.2 Categories

$\mathcal{C}at$ are structures indicating a set of abstract categories, or classes, to which agents entering the system may belong. We assume categories as characterized by a list of *features*, shaped on various dimensions and owned by agents belonging to that category.

- Cat_{Prof} indicates professional and pragmatic abilities, grouping together agents specialized in a given activity. For instance, professional categories refers to *Surgeons, Pediatrist, Oncologists*, etc.
- Cat_{Disp} indicates dispositional abilities, grouping together agents characterized by particular attitudes of willingness in fulfilling their activities. For instance, dispositional categories refers to being *Cautious, Careful, Impulsive* etc.
- Cat_{Cross} indicates crosscutting categories not considered in the above mentioned characterization, for instance being *male, female, young, old*, etc.

[3] The choice of task requirements, features and constraints is arbitrary and aimed at showing the functioning and the efficacy of the categorization reasoning, regardless of the compliance of the real medical domain.

Table 1 (b,c,d) shows examples of categories defined in the medical scenario. Professional and dispositional categories include explicit reference to a range of krypta which one may assume for an agent belonging to that category. We assume agents belonging to a given category as having features in the range specified by that category, for instance a *Pediatrician* agent is supposed to have a *manual* ability between 70 and 100, a *pediatr_spec* between 99 and 100, and so on. On the other hands, crosscutting categories only refers to agent's observable manifesta. As said, krypta can not be automatically inferred from crosscutting categories. Hence, the crosscutting manifesta of being *female* initially has an unknown impact on the task fulfillment. The ability to possibly relate the presence of a given crosscutting manifesta to the effectiveness of the agent in fulfilling the task is up to agent reasoning model (it will be described in the next section).

As can be noticed by matching task requirements and category features, each professional category is shaped by design on the requirements of the specific tasks. In particular we assume at least one specializing feature among the professional abilities of a given category related a given task. For instance, we assume the `Pediatrist` category to be related to the `Chickenpox` task by means of the *pediatr_spec* requirement.

4 Agent Cognitive Architecture

We assume an open MAS where the structure $\mathcal{A}g$ indicates a set of agents, each agent possibly entering and leaving the system at any time, and playing the role patient (trustor) or medical doctor (trustee). We assume patient agents are not able to autonomously fulfill the tasks, thereby they need to delegate its concrete fulfillment to a doctor agent. This section provides a formal description of the cognitive architecture through which agents implements trust based delegation.

4.1 Agent Configuration

We assume each agent $ag_i \in \mathcal{A}g$ represented by the following structures:

$$ag = \langle ag_{attr}, ag_{ep}, ag_{goal}, ag_{cog} \rangle$$

where ag_{attr} a list of agent attributes, ag_{ep} represents agent epistemic states (beliefs), ag_{goal} motivational states (goals), and finally ag_{cog} a set of mechanisms realizing cognitive abilities.

Agent Attributes $ag_{attr} = \langle ag_{id}, ag_{role}, ag_{kr}, ag_{mnf} \rangle$ defines a list of attributes owned by agents:

- ag_{id} is the agent identifier (or agent name);
- ag_{role} defines the role actually played by the agent;
- $ag_{kr} = \langle kr_{Ab}, kr_{Will} \rangle$ defines a set of internal properties (*Krypta*), in particular:

  - $kr_{Ab} = \{\alpha_1, \alpha_2, ...\alpha_O\}$ defines concrete professional abilities to fulfill tasks;
  - $kr_{Will} = \{\omega_1, \omega_2, ...\omega_P\}$ defines concrete dispositional abilities to fulfill tasks;
- $ag_{mnf} = \langle mnf_{Pro}, mnf_{Disp}, mnf_{Cross}\rangle$ defines a list of properties observable by other agents (*Manifesta*), in particular:
  - $mnf_{Pro} = \{\phi_1, \phi_2, ...\phi_Q\}$ refers to signals indicating professional abilities;
  - $mnf_{Disp} = \{\psi_1, \psi_2, ...\psi_R\}$ refers to signals indicating agent's dispositional attitudes
  - $mnf_{Cross} = \{\delta_1, \delta_2, ...\delta_S\}$ refers to signals indicating crosscutting attributes

For instance, professional manifesta may refer to observable signals indicating an agent specialized in pediatrics or in surgery. Dispositional manifesta refers to signals indicating an agent impulsive or cautious. Crosscutting manifesta refers to signals indicating crosscutting categories, i.e., being male or female, etc.

Epistemic States Agent's epistemic states (i.e., beliefs) are represented by the following main structures:

$$ag_{ep} = \langle \mathcal{O}thers, \mathcal{C}at, \mathcal{M}em\rangle$$

$\mathcal{O}thers$ includes an explicit representation for every other agent actually playing inside the system. We assume that an agent ag_i explicitly represent another agent $ag_k \in \mathcal{O}thers$ by storing ag_k's manifesta properties:

$$ag_k = \langle ag_{id}, ag_{mnf}\rangle, \quad ag_k \in \mathcal{O}thers$$

where ag_{id} is the agent identifier, and where ag_{mnf} indicates the signals observed by ag_i upon ag_k.

$\mathcal{C}at = \langle Cat_{Prof}, Cat_{Disp}, Cat_{Cross}\rangle$ indicates the set of categories related respectively to agent professional abilities, dispositions and cross categorial features. In concrete implementation, we assume that the properties observable in a given agent (manifesta) can be automatically retrieved by perceiving the environment. We also assume for the patients a complete knowledge of categories and manifesta in terms of symbolic beliefs.

Finally, $\mathcal{M}em$ builds up the memory of the agent, and it is realized as a belief set storing in patients belief base the results of past delegations.

Motivational States As said, at each round trustor agents (patients) receive a task to fulfill, and for each task they adopt a goal aimed at delegating the activities needed to fulfill it to some trustee (doctors). Such a goal has the following structure:

$$\gamma = \langle \tau, \gamma_{cog}, \gamma_{options}\rangle, \quad \gamma \in ag_{goal}$$

where $\tau \in \mathcal{T}$ is the task associated to that goal, and, from an agent perspective, is given by:

Algorithm 1 Patient delegations process

```
Variables:
  $\tau$ : Task to fulfill.                          $\mathcal{C}at$ : Categories.
  $\mathcal{O}thers$ : Unknown agents.               $\mathcal{M}em$ : Belief set storing results of past delegations.
  $\gamma_{options}$ : Potential trustees.   $task\_cat\_eval$ : Belief set indicating how much a categories fit tasks.

procedure $delegate(\tau)$
  $task\_cat\_eval = \mathsf{ascribe}_\tau(\tau, \mathcal{C}at)$
  $\phi_\tau = \mathsf{fcm}_\tau(\tau)$
  for each $ag_i \in \mathcal{O}thers$ do
      if $\mathsf{matches}_\tau(ag_i, \tau) \neq \perp$ then
          $DoT_{ag_i,\tau} = \mathsf{trust\text{-}eval}(\mathcal{M}em, task\_cat\_eval, \phi_\tau)$
          $\gamma_{options} = \gamma_{options} \cup \langle ag_i, DoT_{ag_i,\tau} \rangle$
      end if
  end for
  $trustee\_agent = findBest(\gamma_{options})$
  send($trustee\_agent$, achieve, $\tau$)
procedure $response(Trustee, \tau, Result)$
  $\mathcal{M}em = \mathcal{M}em \cup \langle Trustee, \tau, Result \rangle$
```

- $\tau_{Prof} = \{\alpha_1, \alpha_2, ...\alpha_O\}$ describes the abilities needed to fulfill the task;
- $\tau_{Disp} = \{\omega_1, \omega_2, ...\omega_P\}$ describes the willingness (dispositions) needed to fulfill the task

Notice that agents ignore τ_{Cross}. In fact, we are assuming a lack of causal knowledge—thus agents which initially are not able to understand how cross categorial features may influence the task. γ_{cog} is the particular cognitive module which is configured to decide to which other agent delegate the task. As will be shown in the next sections, in concrete implementation γ_{cog} is realized through a Fuzzy Cognitive Map (FCM). Finally, $\gamma_{options}$ is a list of possible trustees selected for the delegation. In this case, it represents the options to delegate the task to the trustees. Each element in $\gamma_{options}$ is of the form: $\langle ag_{id}, trust_{id} \rangle$, where ag_{id} indicates a trustee identifier, and $trust_{id}$ represents its related trust value (with $-1 \leq t \leq 1$).

Cognitive Modules In order to find a list of potential trustees for a given task, the trustor has to assess a value of trust each of them. The abstract specification of the trust evaluation model is shown in Alg. 1. It uses a series of cognitive mechanisms and heuristics defined inside ag_{cog}. In particular, ag_{cog} are elements of the type $\langle \Phi, \Psi \rangle$, where Φ represents a decisional module (realized through a Fuzzy Cognitive Map-FCM and described in the next section), and where Ψ includes a set of reasoning abilities, resumed by: (i) $\mathsf{ascribe}_\tau$, (ii) $\mathsf{matches}_\tau$, (iii) fcm_τ, (iv) $\mathsf{trust\text{-}eval}_\tau$.

The $\mathsf{ascribe}_\tau$ function, given the specification defined for one task and for each category, allows to quantify the relationship between each category and the specified task:

Definition ($\mathsf{ascribe}_\tau$ - *Associating a Task to Categories*) Let be the representation for a given goal adopted by an agent $\gamma = \langle \tau, \gamma_{cog}, \gamma_{options} \rangle$. Let $\mathcal{C}at \in ag_{ep}$ a belief set indicating professional and dispositional categories. We define: $ascribe_\tau : \mathcal{T} \times \mathcal{C}at \rightarrow ag_{ep}$ as the function $\in \Psi$ finding a series of ex-

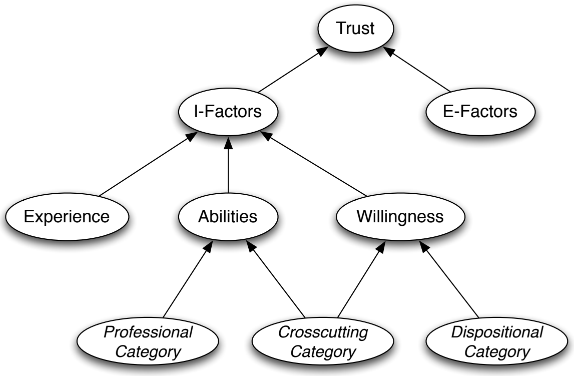


Fig. 1. FCM used by trustor agents to assess the degree of trust of possible trustees.

pressions indicating the matchmaking value between category constraints and the task requirements. In other terms, given the representation of a given task τ, $\mathsf{ascribe}_{\tau,\mathcal{C}at}$ retrieves to which extent the task τ matches the categories $\in \mathcal{C}at$. In concrete implementation, this function produces a set of beliefs to be stored in ag_{ep} relating the task τ to the elements in Cat_{Prof} and Cat_{Disp}. In Alg. 1 (row 1), such a beliefs have the form: `task-cat-eval(Task, Category, ascribe(Task, Category)`.

The $\mathsf{matches}_\tau$ function allows to quantify how a potential trustee belonging to a given category has the required features to fulfill the task or not:

Definition ($\mathsf{matches}_\tau$ - *Matching agent Abilities and task Requirements*) Let $ag_{mnf} = \langle mnf_{Pro}, mnf_{Disp}, mnf_{Cross}\rangle$ the observable properties for an agent $\in \mathcal{O}thers$. Let $\tau \in \mathcal{T}$ a task including a list of agent abilities and dispositions required to fulfill that task. We define: $\mathsf{matches}_\tau : \mathcal{O}thers \times \mathcal{T} \rightarrow \{1, \perp\}$ as the function $\in \Psi$ returning 1 if the categories required for fulfilling the task match the agent properties, $\perp$ elsewhere. In Alg. 1 (row 4), $\mathsf{matches}_\tau(ag_i, \tau)$ is used to verify whether ag_i, according to its manifesta, is matching the requirements needed to fulfill τ.

Given the requirements defined by each $\tau \in \mathcal{T}$, the fcm_τ function allows to configure the appropriate cognitive architecture for that task:

Definition (fcm_τ - *Modulating Architectures for Tasks*) Let the representation for a given goal adopted by the agent $\gamma = \langle \tau, \gamma_{cog}, \gamma_{options}\rangle$. We define: $\mathsf{fcm}_\tau : \mathcal{T} \rightarrow \Phi$ as the function $\in \Psi$ configuring the cognitive map ϕ_τ suitable for evaluating all the possible trustees to which τ could be delegated. In Alg. 1 (row 2), $\mathsf{fcm}(\tau)$ configures a FCM ϕ_τ to be used by the agent to find the best trustee.

Given the extent according to which categories match the task τ, and given a cognitive map which is configured with respect to τ, the $\mathsf{trust\text{-}eval}_\tau$ function calculates the trust value for any potential trustee in $\mathcal{A}g$. The output of this function indicates a number resuming the trust value actually assessed for a given trustee.

Definition ($\mathsf{trust\text{-}eval}$ - *Associating trust to a trustee*) Let the representation for a given goal adopted by an agent $\gamma = \langle \tau, \gamma_{cog}, \gamma_{options}\rangle$. Let ag_{ep} the belief base including the set $\texttt{task\_cat\_eval}$, matching the task τ with the available categories, and the set $\mathcal{M}em$, as the memory of past delegations. Let $\phi_\tau \in \Phi$ the cognitive map configured for the task τ. Then, we define: $\textsf{trust-eval}_\tau : \mathcal{O}thers \times \Phi \to [-1;1]$ as the function $\in \Psi$ calculating the *trust* value for a given trustee.

In Alg. 1 (row 5), $\textsf{trust-eval}(\mathcal{M}em, task\_cat\_eval, \phi_\tau)$ is applied to each possible trustee in $\mathcal{O}thers$ in order to assess its related trust value.

4.2 FCM Trust Attribution

As said, the mechanism underlining trust-eval is realized through a Fuzzy Cognitive Map (FCM) which is configured on the fly by the trustor agent, given the cognitive module $\textsf{fcm} \in ag_{cog}$ described above. FCMs allow for a flexible computational design of the cognitive model described in Section 2, making it available a straightforward decision making function in different applications and domains [10, 6]. Cognitive maps models a causal process by identifying a series of *concepts* and *causal relations*, being represented as a *weighted graph.* The functioning is governed by Fuzzy Logics [11]: at each computation step, the value of a concept is updated by calculating the impact provided by the other concepts (i.e., the weighted sum of the fuzzy values of the incoming edges). Such a value is squeezed from a specified node's activation function and the computation continues until a convergence is reached.

Fig. 1 shows the FCM used inside the trust-eval mechanism. It is a tree-like structure having *Trust* as root concept. The two main contributions to trust are *external* and *internal factors.* The i-factors are the elements depending on the internal characterization of the trustee, i.e given by trustee's internal capabilities to fulfill the specified task. This node is attached to the two sub-nodes resuming trustee's *abilities* and *willingness.* Each of these nodes is linked to the professional and dispositional categories defined for this domain (see Table 1). The weight of the link reflects the *impact* of the category on the task, as it is computed by the function $\textsf{ascribe} \in ag_{cog}$.

The adopted FCM uses *identity activation function* and is built so as trust values converge within the interval [-1,1] and no approximation errors is propagated by squeezing the values. We mean the negative subinterval [-1,0] as *mistrust*, namely the case when agent distrusts from delegating the task to another agent. The value 0 means *neutral trust* or absence of trust at all.

This template of the map allows for different types of cognitive evaluations of trust by inactivating or pruning some branches. Indeed, in the special case where also direct experience is considered, a further leaf node *"experience"* is attached to the internal factors. In the scenario discussed in this paper, the trustor uses only *i-factors* branches (related to manifesta and ascribed categories), thus the *e-factors* branches can be excluded from the computation. Instead, *e-factors* branches can be activated for those agents able to understand how the environmental conditions are going to affect the trustee performance.

The concrete implementation of the Alg. 1 is realized as an hybrid architecture. The fuzzy modules through which the cognitive maps are managed is added on top of a BDI engine. The open source project Cog-Trust is used

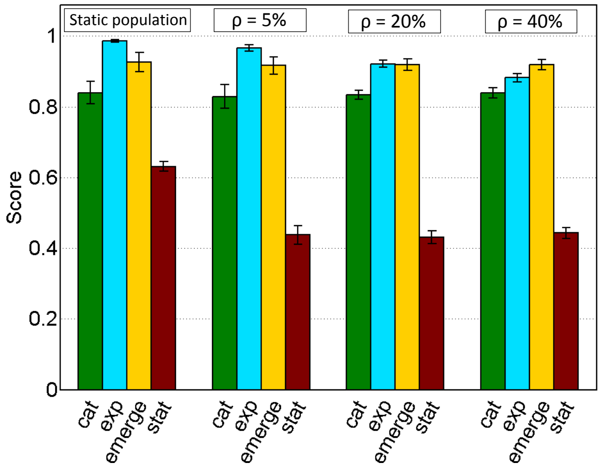


Fig. 2. Mean scores achieved by trustor agents engaged with the task *chickenpox*, in varying conditions.

to implement the cognitive modules, while the BDI engine is realized using the Jason platform [3]. The Jason communication infrastructure is used to realize a simplified contract-net between trustor and trustee agents[4].

5 Experiments

This section presents the experimental evaluation for agents in repeated trials. Experiments observe how different trust formation strategies affect the individual performances of the agents in evolving experimental conditions. Each experiment consists of R *rounds* at the beginning of which, every trustor receives a specific task from the simulator engine. Trustor's goal is to find the best trustee to delegate the task among a population of N possible trustees. An heterogeneous set of trust formation strategies is analyzed. In detail, the following six delegation strategies are considered:
Cat. This strategy is based on the cognitive architecture realizing the categorial reasoning described in Section 4. Categorizing agents are thus able to prune the set of possible trustees looking for those categories that guarantee the best expected outcome. Trust values are computed using a FCM (Fig. 1) including the nodes of internal factors related to abilities and willingness. The map is built according to what said in the previous section and it is populated with the manifesta properties of the trustee retrieved from $\mathcal{M}em$. The FCM mechanism assigns a higher trust value to the trustees who belongs to the professional and dispositional categories better fitting the task requirement. The connections between perceived manifesta and internal FCM nodes are established by the ascribe function, measuring the features matching on the ongoing task.

[4] The CogTrust architecture, along with the experiments described in this paper, are available as an open source project at `mindraces-bdi.sf.net`.

Exp. Experience agents add to the FCM used by Cat a further branch summarizing the personal knowledge of the evaluated trustee. Past experiences are resumed for each trustee for the given task. The leaf of the personal experience branch is filled with the values coming from the average of the previous individual performances, as they have been stored in $\mathcal{M}em$.

Stat. The statistic agent uses only personal knowledge. It finds the best trustee on the basis of the history of the previous interactions stored in $\mathcal{M}em$. At each task completion, Stat stores the result value of task fulfillment by the delegated trustee to be used as a index of trust in the next encounters with the same agent.

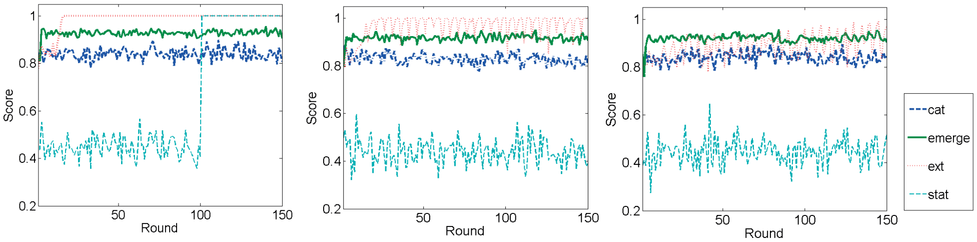
Emerge. Emerge agents combine categorial and personal reasoning in order to dynamically refine and adjust the trust-eval mechanism used by Cat. Information about crosscutting manifesta is exploited in order to let to *emerge* a set of abstract categories related to the encountered crosscutting manifesta (i.e., being male, female, etc.). Such a crosscutting categories have not a direct relation with abilities and willingness as in the case of professional and dispositional ones, although they concretely influence the performance of the trustee. In order to learn how the emergent category affect trustee's performances, Emerge agents apply a learning mechanisms as part of their trust-eval mechanism. In particular, Emerge agents build clusters inside $\mathcal{M}em$ grouped by crosscutting categories. On such a basis, they try to update the $task\_cat\_eval$ related to the crosscutting categories based on their personal experience.

Fulfillments are measured by absolute scores, referred as the fraction of the highest performance value reachable in the current population for the given task. At the initialization, the simulation engine selects randomly 100 trustees from a repository of 2500 predefined profiles with a random distribution of categories, krypta and manifesta. Openness is measured in terms of population changes. The number of rounds in which the population is fixed forms a *Era*. At the end of each *Era*, ρ% of the trustee population is replaced by new trustees. In the current setting we use $Era = 5$ rounds. Each experiment is characterized by the score trends averaged for 20 simulations. For simplicity, the experiments have a fixed task (`Chickenpox`), for which the fulfill function speculates that females perform 10% better than males. Experiments have been run on a machine Inter(R) Core(TM) i5 CPU x64, 2.67 MHz, 6GB RAM, and using Jason 1.3.

5.1 Results

Experiments analyzed how trustor's performance is affected by the frequency and the size of the changes inside the population. We first analyzed agents dealing with a static population and then we progressively increased the ρ parameter to see the effects on the delegation when a small, medium and large part of the population changes. In what follows, we discuss the results for $\rho = 0$, $\rho = 5$, $\rho = 20$ and $\rho = 40$ (Fig. 2).

Fixed Population. Fixed population hypotheses observes trust formation when the population is static (no trustee replacements and $\rho = 0$). In this case direct experiences result a relevant source of information for trust formation. The Exp agent turns to be the best delegator. Being able to exploit the categorization



(a) $\rho = 0$ (b) $\rho = 5$ (c) $\rho = 40$

Fig. 3. Evolution of the trustor scores in any rounds for the task *chickenpox*, varying the ρ parameter, with $Era = 5$ rounds.

reasoning joint to the experience of past delegations, it gets the optimal delegation strategy finding the best trustee within the population (Fig. 2). Stat gets a lower ranking, although its score would be the same of Exp excluding the learning phase spent during the first 100 iterations.

Thanks to the cognitive attribution of trust using categorization and FCM based trust_eval, the exploration of the cognitive agents Cat, Emerge, and Exp is limited to the only specialized trustees (Pediatricians) for the current task. They prune the search space thus wasting less time to find the best performer than the Stat agent. Cognitive attribution of trust based on personal and categorial reasoning allow to quickly stabilize delegation outcomes on the maximum value. The advantage in score of 10% for the Emerge, compared to the Cat agent, is due to the categorial reasoning that let to emerge a preference for females.

Open Populations. Open population hypotheses assume that trustees can leave and can be replaced by others during the simulation. This dramatically increases the probability to face new unknown trustees. Accordingly, openness strongly influences the effectiveness of reasoning on the personal level through direct experiences stored in memory.

When $\rho = 5$, Stat agents show random delegation choices as they are forced to continuously test all the new incoming trustees (Fig. 3(b,c)). The increase of ρ also narrows the gap between Exp agent and the two others categorizer agent: Emerge and Cat. In fact, Fig. 3(b) shows the occurrence of many low scores in the Exp's profile due to the fact that it is not able to further refine the crosscutting categories. $\rho = 20$ is the balance-point, in which Exp and Emerge equalize their scores on 0.93 (Fig. 5, mid-right). For $\rho \geq 40$, Exp finally loses his advantage, as the large replacement of doctor trustees obliges it to compute a new search for the best. Exp totally gets a score of 0.87 while Emerge is the winner with 0.93.

5.2 Discussion

As results point out, agents reasoning on the personal level need to explore the whole population to find the best performer, thus requiring a huge amount of time and resources before reaching an effective result. On the contrary, the combination of categorial reasoning and direct experience promotes an effective exploration strategy. Results confirm that categorial trust is robust to any population change: Cat and Emerge keep the same scores, regardless of the variation of the ρ parameter. The good results of categorizer agents is supported by the computationally efficient implementation of the categorial experience, using the search space $O(|Cat|)$, against $O(|Ag|)$ space required for the individual experience.

Thanks to the FCM structure adopted for trust formation, the distinctive feature of the cognitive trustors is the ability to combine three levels of reasoning: (i) the *categorial level* considers abilities and dispositions of the trustee seen as a member of a known class or category; (ii) the *personal level* is concerned with the direct experiences; (iii) the *environmental/contextual level* which is is concerned with the situation influencing the performances in specific contexts. Facing openness and dynamic populations complicates the delegation, as repeated interactions with the same agent are rare and direct experience mechanisms become increasingly unreliable. This context emphasizes trustor's ability to refine and revise categories, forming *general correlations* and *evaluations* based on the interaction with individuals. Categorization is a twofold reasoning process. Assuming an agent in a class or category is a form of *generalization* from single experiences to form general correlations and evaluations. On the other side, this also allow to transfer, "instantiate", the attributes and features of that general class on a given individual agent.

6 Conclusions

This work describes and evaluates a cognitive architecture based on a model of trust for agents able to reason in terms of categories, against the current approaches which are mostly based on the personal level (reputation, direct experience, observation and statistical analysis). This approach provides an alternative approach to dynamic and open systems. Experimental analysis showed that delegation effectiveness does not depend on the composition of the population, but the model is resistant to mutations and replacements, and it also benefits of efficiency of having reduced categorial information instead of extensive individual experiences.

Limitation of the current approach pave the way to future works. At an architectural level, a seamless integration between the deliberative and cognitive modules will be be studied. The computational model actually forces the developer to specify a FCM template, and then to tune its functioning through an off-line setting of weights and connections. Future work will account the ability of agent to learn connections and adapt the functioning of their cognitive modules on the fly. Another drawback is the need for agents to know a pre-established set of categories ($\mathcal{C}at$). Further studies will explore agents *unifying* personal and categorial level, i.e. autonomously creating new categories from scratch on the basis of individual experiences.

References

1. Michael Bacharach and Diego Gambetta. Trust as Type Detection. In *Trust and deception in virtual societies*, 2001.
2. B. Barber. *Logic and the limits of Trust.* Rutgers University Press, 1983.
3. Rafael H. Bordini, Jomi Fred Hübner, and Michael Wooldrige. *Programming Multi-Agent Systems in AgentSpeak using Jason.* Wiley Series in Agent Technology. John Wiley & Sons, 2007.
4. C. Burnett, T.J. Norman, and K. Sycara. Bootstrapping Trust Evaluations through Stereotypes. In *Int. Conf. on Autonomous Agents and Multiagent Systems (AA-MAS 2010)*, pages 241–248, 2010.
5. Cristiano Castelfranchi and Rino Falcone. *Trust Theory. A Socio-Cognitive and Computational Model.* John Wiley & Sons, 2010.
6. R. Falcone, G. Pezzulo, and C. Castelfranchi. A fuzzy approach to a belief-based trust computation. *Trust, reputation, and security: theories and practice*, pages 55–60, 2003.
7. Karen K. Fullam and K. Suzanne Barber. Dynamically learning sources of trust information: experience vs. reputation. In *Int. joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-07)*, pages 164:1–164:8, 2007.
8. J.F. H ubner, E. Lorini, L. Vercouter, and A. Herzig. From cognitive trust theories to computational trust. In *Workshop On Trust in Agent Societies (Trust@AAMAS09)*, 2009.
9. T. G. Huynh, N. R. Jennings, and N. R. Shadbolt. An integrated Trust and Reputation model for Open Multi-Agent Systems. *Journal of Autonomous Agent and Multi-Agent Systems*, 13:119–154, 2006.
10. B. Kosko. Fuzzy Cognitive Maps. *International Journal of Man-Machine Studies*, 24(1):65–75, 1986.
11. B. Kosko and J.C. Burgess. Neural Networks and Fuzzy Systems. *The Journal of the Acoustical Society of America*, 103:3131, 1998.
12. Michael L. Littman and Peter Stone. Leading Best-Response Strategies in Repeated Games. In *IJCAI 2001 Workshop on Economic Agents, Models, and Mechanisms*, 2001.
13. Jordi Sabater-Mir, Mario Paolucci, and Rosaria Conte. Repage: Reputation and image among limited autonomous partners. *Journal of Artificial Societies and Social Simulation*, 9(2), 2006.
14. M. Wojcik, J. Eloff, and H. Venter. Trust model architecture: Defining prejudice by learning. In *Trust and Privacy in Digital Business*, volume 4083 of *Lecture Notes in Computer Science*, pages 182–191. Springer, 2006.

A Robust Collective Classification Approach to Trust Evaluation

Xi Wang, Mahsa Maghami, and Gita Sukthankar

Department of EECS
University of Central Florida
Orlando, FL
{xiwang,gitars}@eecs.ucf.edu,mmaghami@cs.ucf.edu

Abstract. In this paper, we present a collective classification approach for identifying untrustworthy individuals in multi-agent communities from a combination of observable features and network connections. Under the assumption that data are organized as independent and identically distributed (i.i.d.) samples, traditional classification is typically performed on each object independently, without considering the underlying network connecting the instances. In collective classification, a set of relational features, based on the connections between instances, is used to augment the feature vector used in classification. This approach can perform particularly well when the underlying data exhibits homophily, a propensity for similar items to be connected. We suggest that in many cases human communities exhibit homophily in trust levels since shared attitudes toward trust can facilitate the formation and maintenance of bonds, in the same way that other types of shared beliefs and value systems do. Hence, knowledge of an agent's connections provides a valuable cue that can assist in the identification of untrustworthy individuals who are misrepresenting themselves by modifying their observable information. This paper presents results that demonstrate that our proposed trust evaluation method is robust in cases where a large percentage of the individuals present misleading information.

Keywords: collective classification, homophily, agent reputation and trust

1 Introduction

Deciding whom to trust in the absence of direct transactional history is a difficult problem [35] for an individual agent interacting with an open system of self-interested agents. One oft-used mechanism is the direct solicitation of reputation information from a trusted source [36, 45], or multiple, less-reliable sources [13], to avoid deceptions perpetrated by groups of colluding agents. Yet, what if it is not possible to directly query an agent's reputation, either due to communication constraints or a lack of willingness from an agent's fellows to directly testify about past transactions? Here, we suggest that the structure of the network implicitly bears witness to the trustworthiness of the connected agents, regardless of whether the agents directly volunteer reputation information. Our collective classification framework for trust evaluation leverages a combination of observable features and network connectivity to improve performance over non-relational classification paradigms, in addition to making the trust evaluation process more robust against the deceptive efforts of untrustworthy agents.

In this paper, we describe collective classification and show how it can be used for general trust evaluation problems such as coalition building in social networks. Section 2 provides an overview of the social forces driving the creation of human networks and describe how the end result of these forces is an increase in the informative power of the network. Section 3 describes our specific agent reputation and trust scenario in which an individual agent has to evaluate the trustworthiness of a large number of surrounding agents. To make the collective classification process tractable for large datasets we use a local algorithm (Iterative Classification Algorithm, summarized in Section 5). We demonstrate that our framework is highly robust to deceptive agents and generalizes to trust evaluation scenarios in many types of networks. Section 6 presents results on the effects of network factors such as homophily and degree, the use of different types of relational features, and robustness to increasing amounts of deception. In Section 7 we discuss related work in the area and conclude in Section 8.

2 Informative Networks

Network structure can be intrinsically informative when social forces affect the probability of link formation. Human networks often possess the property of homophily, an increased propensity for like-minded individuals to be connected, colloquially described with the phrase “birds of a feather flock together” [27]. Homophily in trust levels could be categorized as a form of value homophily, the tendency of humans to preferentially connect with people who share the same attitudes and beliefs. Along with value homophily, status homophily, preferential linkages created on the basis of attributes such as age, gender, or ethnicity, is commonly observed in human social networks [20]. Network research has shown that the homophily principle creates strong interpersonal network ties in a wide variety of contexts (e.g., neighborhoods, communities, schools) and affects the choice of informal trusted contacts selected for advice and social support [42]. Clearly, since it is often beneficial for deceptive agents to maintain connections with a network of “dupes”, heterophily in trust levels (connections to dissimilar agents) will also exist in trust networks.

A second factor affecting the probability of link maintenance is the agents’ satisfaction with past transactions. In most situations, it is reasonable to assume that agents will preferentially maintain connections with trustworthy agents since those relationships are likely to result in direct benefits [35]. Additionally, agents will form and maintain relationships of convenience driven by factors such as proximity, interaction costs, and supply/demand constraints that are not simply explained by either link prediction model [14]. Regardless of these additional factors, we believe that the network structure remains an informative source of information when either value homophily or transactional satisfaction affect link formation.

An underlying assumption of traditional classification methods is that the instances are independent of each other. On the other hand, networks of agents contain instances of multiple types that are related to each other through different types of links. To classify, or label the node in the network, three classification methodologies have been studied over the last decade. Traditional classifiers, often referred to as the content-only classifier, ignore the network and utilize attribute dependencies to predict the label of unknown instances. Relational classifiers improve classification performance by taking advantage of dependencies of both attribute and labels between related labeled instances [24]. Finally, collective classification aims to simultaneously classify related instances to determine the label of the test node [31, 25]. Studies in other domains have shown that collective classification can increase classification accuracies over non-collective methods when instances are interrelated [30, 41, 22, 47].

3 Problem Formulation

Consider the following scenario. An individual agent in a large, open multi-agent system would like to create the largest possible coalition of trustworthy agents for a joint venture. The agent can access the following information:

1. observable features correlated with the agents' trustworthiness;
2. the existence of links connecting agents that have a history of past transactions (but without weights or valences denoting the outcome of the transactions);
3. a set of labels containing information about the trustworthiness of select members of the community.

Note that each link is meant to serve as summary of past transactions rather than representing the outcome of a single transaction. The agent forming the coalition cannot take any probing actions before making its decision. It is assumed that deceptive agents in the system attempt to foil the trust evaluation by two mechanisms:

1. emitting deceptive features;
2. modifying their labels to appear more trustworthy.

For verisimilitude, the network is assumed to follow a power law degree distribution like many human networks, and link formation is driven by a combination of value homophily, transactional satisfaction, and randomness. As a result, there exists a society of N agents connected by graph G. In this graph the set of nodes, $V = \{V_1, \ldots, V_n\}$, represents the agents; agents are connected by directed links based on the underlying interactions between the agents. The agents' behavior during interactions is modulated by their own internal value system or *trustworthiness*. The true level of an agent's trustworthiness is hidden from the other agents and can assume a label from the set $L = \{L_1, \ldots, L_n\}$.

Each agent i, has two types of attributes: 1) a static feature vector, $S_i = \{s_1, \ldots, s_m\}$, of length m; and 2) a dynamic or relational feature vector, $R_i = \{r_1, \ldots, r_n\}$, of length n. The static feature vector is observable to all the agents and is related to the agent's trustworthiness; example features could include properties such as "returns library books", "answers email promptly", or "reciprocates invitations". Dynamic, relational features, are calculated through aggregating any known labels of connected agents. The set of agents, N, is further divided into two sets of agents: X, the agents for whom we know labels (acquaintances or people known by reputation), and Y, the agents whose label or trust level need to be determined (strangers). Our task is to determine the labels of the unknown agents, Y, from the label set L, based on their two types of attributes. The ultimate goal of the observing agent is to recognize the trustworthiness of other agents in the graph and to form a coalition consisting of the most trustworthy set of agents.

4 Agent Network Generation

To evaluate the performance of collective classification on identifying agents' trustworthiness in a variety of networks, we simulate the evolution of agent networks formed by the combined forces of value homophily and transactional satisfaction. Since social communities often form a scale-free network, whose degree distribution follows a power law [1], we model our agent networks in the same fashion.

Following the Sen et al. [38] network data generation method, we control the link density of the network using a parameter, *ld*, and value homophily between agents using a parameter, *dh*. The effects of value homophily is simulated as follows:

1. At each step, a link is either added between two existing nodes or a new node is created based on the link density parameter (*ld*). In general, linking existing nodes results in a higher average degree than adding a new node.
2. To add a link, we first randomly select a node as the source node, A, and a sink node, B, based on the homophily value (*dh*), which governs the propensity of nodes with similar trustworthiness values to link. Node B is selected among all the candidate nodes in the correct class, based on the degree of the node. Nodes with higher degree have a higher chance to be selected.

Transactional satisfaction also governs the process of link formation. Once the link generation process starts, we add a directed link from node A to node B by default, under the assumption that the first selected agent initiated the transaction. The transactional trustworthiness of the second node governs whether a reciprocal link is formed. Here, we use an evaluation function $F_x(p, t)$ to map an observed performance value p in a particular task t to a binary evaluation of performance (positive or negative). We assume that all agents use the same evaluation function for all tasks, which is:

$$F_x(p,t) = \begin{cases} 1: & p \geq 0.5 \\ -1: & p < 0.5 \end{cases}$$

To generate a new node, we first select a trustworthiness level based on a uniform class distribution and assign that class label to the node. Then we add links between the new node and one of the existing nodes as we described above. Inspired by the model proposed by Burnett et al. [5], the trustworthiness label (Table 1(b)) governs the mean and standard deviation parameters of a Gaussian distribution from which simulated performance values are drawn. The algorithm for simulating the evolution of the agent network is outlined in Table 1(a).

After generating the network, we assign observable static attributes to each agent by drawing from a set of binomial distributions based on its trustworthiness. Attributes are represented as a binary feature vector, which indicates the existence or absence of a given feature. These features are meant to represent observable properties that result from the consistent practice of an agent's trust value system. Observable attributes for each class are generated using a set of binomial distributions. Attributes are represented by a binary feature vector, length 10, but the maximum number of attributes that can be true is capped at 5. Random noise is introduced to the attribute generation process using the *attrNoise* parameter. Specifically, with a probability of *attrNoise*, each binary feature is independently assumed to be corrupted, in which case it is set randomly to either 0 or 1 with equal probability. The *attrNoise* parameter can be used to model the

Table 1.

(a) Agent Network Generator

```
Agent Network Generator (numNodes, ld, numLabels, attrNoise, dh)
$i = 0$
G = NULL
while $i <$ numNodes do
    sample $r$ from uniform distribution $U(0, 1)$
    if $r \leq ld$ then
        connectNode(G,numLabels,dh)
    else
        addNodes(G,numLabels,dh)
        $i = i + 1$
    end if
end while
for $i$ = 1 to numNodes do
    Attributes = genAttr($v$, Attributes, label, attrNoise)
    where $v$ is $i$th node in G
end for
return G
```

(b) Agent Task Performance Profile

Trust Level	Mean	StDev
L1	0.9	0.05
L2	0.6	0.1
L3	0.4	0.1
L4	0.2	0.05

level of deceptiveness of agents in attempting to hide observable attributes that provide clues about their trustworthiness.

5 Iterative Classification Algorithm

In this agent network scenario, collective classification refers to the combined classification of a set of interlinked nodes using three types of correlations [39]: 1) correlations between the label of node V and its observed attributes; 2) correlations between the label of node V and the observed attributes (including observed labels of nodes in its neighborhood); and 3) the correlations between the label of node V and the unobserved labels of agents in its neighborhood. For our experiments, we use the iterative classification algorithm [30], an approximate inference algorithm that has shown promise at hyperlink document classification tasks.

Iterative classification was first proposed by [30] and has since been extended by [26]. In ICA, the training model is built using both static and relational attributes of the observed nodes. Since the class labels of the training nodes are known, the value of the dynamic attributes can be calculated using aggregation operators such as *count*, *proportion*, or *mode*. Aggregation operators are different ways of representing the same information (the labels of the connected nodes), but alternate representations have been shown to impact classification accuracy, based on the application domain [39].

The training model is applied to the test nodes whose class labels are unknown; in our problem, these are the stranger agents, for whom no reputation information exists. Initially, because some of class labels of the related nodes are unknown, the values of their relational attributes are also unknown. This problem can be solved by bootstrapping the classification process. At the beginning, the prediction of the class labels for all test nodes is obtained using content features only. Predictions made with high probability are accepted as valid and are accepted into data as known class labels. After certain percentage of classification with highest probability are accepted, the classifier recalculates the relational attributes using the newly accepted labels and reclassifies the labels. In each iteration, a greater percentage of classifications are accepted and new dynamic attributes are filled in. It is worth noting that the prediction is both recalculated and reevaluated in each iteration; hence the prediction about a given node might change over the process of iteration. Therefore, the label of a node accepted in one iteration might be discarded in the next iteration if the probability associated with the prediction is no longer in the top percentage of acceptance predictions. ICA has the potential to subsequently improve classification accuracy on related data after iterations. However, it should be carefully applied since the incorrect relational features in one iteration may diminish the classification accuracy. Table 2(a) shows the pseudo-code for ICA.

Experiments have shown improvement in classification accuracy by making certain modifications to basic ICA. For instance, [26] proposes a strategy where only a subset of the unobserved variables are utilized as inputs for feature construction. More specifically, in each iteration, they choose the top K most confident predicted labels and use only those unobserved variables in the following iterations predictions, thus ignoring the less confident predicted labels. In each subsequent iteration they increase the value of K so that in the last iteration all nodes are used for prediction.

In this paper, we explore the use of a reputation-based aggregation operator. For a rational agent, its reputation in a trust system is often calculated based on evidence consisting of its observable positive and negative experiences [43]. This evidence can be collected by an agent locally or via a reputation agency. We define the agent's reputation as the average judgment by its observable direct interactions. We assume that the agent will receive a positive evaluation only if its interactors's trust level is equal or lower than itself's. The agent's reputation is therefore the frequency of positive opinions. Suppose $r_x^{N_x}$ is the number of positive evaluation agent x received from its observable interactors N_x, and $s_x^{N_x}$ is the number of negative evaluations. We compute reputation based on $r_x^{N_x}$ and $s_x^{N_x}$ as

$$R_x = \frac{r_x^{N_x}}{r_x^{N_x} + s_x^{N_x}}. \tag{1}$$

Note that R_x is a single scalar value, unlike typical aggregation operators such as *count* or *mode*.

6 Experiments

Our experimental methodology can be summarized as follows. We generate agent networks using the procedure described in Section 4 with the network parameter values specified in Table 2(b). *numNodes* refers to the total number of agents in the network, including both agents whose trust levels are revealed (analogous to the training set) and those for which trust levels are hidden (corresponding to a test set); *dh* denotes the homophily of the network; *numLabels* is the number of discrete trust levels, with 1 corresponding to the most trustworthy agents; *numFeatures* is the dimensionality of the binary feature vector; *attrNoise* controls the probability that a given binary feature is randomized (corresponding to a degree of deception). Unless indicated otherwise, these parameter values are fixed across experiments and plot classification accuracy against the link density of generated networks. For each network instance, we perform three-fold cross-validation (using disjoint subsets of agents with revealed and hidden labels) and report averaged results.

To evaluate the performance of collective classification in defining the trust level of unknown agents, we adopt the ICA algorithm [26] and employ the Logistic Regression Classifier (LRC) as the baseline classifier in all the experiments.

Table 2.

(a) ICA

```
Iterative Classification Algorithm
1. Build model on fully labeled training set.
2. Apply trained model to test set of $N$ instances.
   For each iteration $i$ : 1 to $K$
   a. Calculate values for dynamic relational attributes
   b. Use model to predict class labels
   c. Sort inferences by probability
   d. Accept $m$ class labels, where $m = N \times (i/K)$
3. Output final inferences made by model on test set
```

(b) Parameter settings

Parameter Name	Value
numNodes	500
dh	0.8
numLabels	4
numFeatures	10
attrNoise	0.2

We perform a series of experiments to investigate several key issues in collective classification for trust evaluation. First, we compare collective classification against a baseline classifier, both in terms of overall accuracy and on inter-class misclassification. We then explore how the benefits of collective classification depend on network characteristics, such as link density and homophily. We also evaluate the impact of a variety of aggregate operators that represent the relations between trust levels of connected agents and finally examine the robustness of collective classification to two forms of deception in agent networks.

6.1 Comparisons against baseline classifier

Figure 1 compares the classification accuracy of ICA against the baseline classifier (logistic regression) for default agent network parameter settings. The feature vector for the baseline algorithm is simply the list of observable binary features, while that of ICA is augmented by the agent's relational attributes expressed using the *count* operator. The latter is a histogram over trust levels of the agents connected to the given agent, computed in both directions (i.e., an additional 8-dimensional feature). As can be seen from the graph, ICA improves over the baseline in a small number of iterations and converges rapidly. Based on this, we use the same value of $K = 10$ for the number of ICA iterations. More importantly, we observe that ICA dramatically improves the classification accuracy from a baseline of 73% to 95%, showing that collective classification is able to exploit significant information about agent trust levels encoded in the network, beyond that expressed in the observable features alone.

Tables 3 presents the confusion matrices for the baseline (LRC) and collective classification (ICA) approaches. We can make several observations about the misclassifications. First, collective classification virtually eliminates the possibility of misclassifying an agent as very untrustworthy (L4). Second, the classification accuracy for L1–3 agents improves dramatically. Finally, although the classification accuracy of L4 agents remains unchanged, we see that ICA is much less likely to misclassify L4 agents as trustworthy (L1).

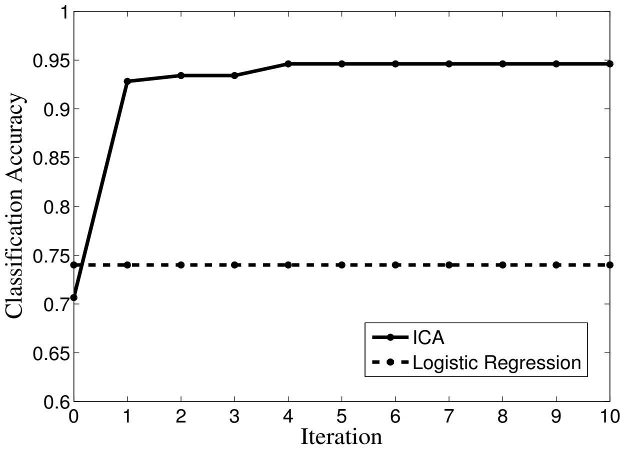


Fig. 1. Collective classification (ICA) clearly outperforms the baseline (LRC) and converges in a few iterations. (*ld*=0.4, *dh*=0.8,*attrNoise*=0.2).

Table 3. Confusion matrix for baseline (on left) and collective classification (on right) with parameter setting *ld*=0.4,*dh*=0.8,*attrNoise*=0.2

	L1	***L2***	***L3***	***L4***
L1	80.0	14.3	5.7	0
L2	16.3	60.5	20.9	2.3
L3	2.6	5.1	74.4	17.9
L4	6.0	0	6.0	88.0

	L1	***L2***	***L3***	***L4***
L1	97.1	0	2.9	0
L2	9.3	90.7	0	0
L3	0	2.6	97.4	0
L4	2.0	2.0	8.0	88.0

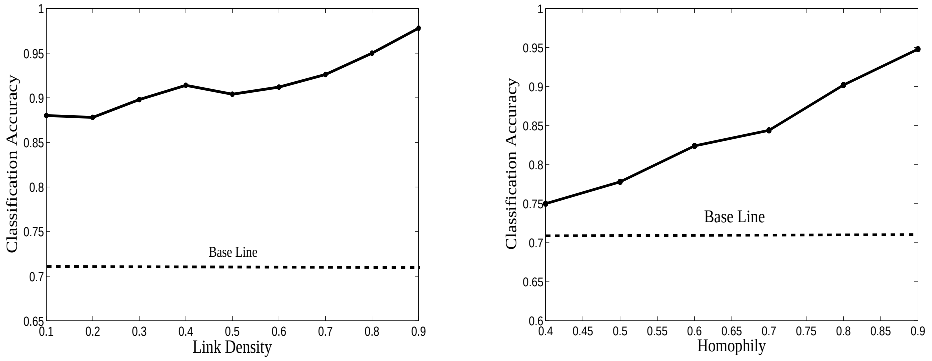
6.2 Link density and Homophily

In order to observe the impact of network's link density parameter on collective classification, we generate networks with *ld* changing from 0.1 to 0.9 with step size 0.1, and freezing *attrNoise* and *dh* at 0.2 and 0.8, respectively. Figure 2(b) shows how ICA classification accuracy varies with link density. The results show that ICA continues to outperform the baseline and that classification accuracy improves with increased link density. These results are consistent with our expectation that where reliable dependencies exist between instances, increasing the degree of links enables collective classification to more reliably extract relational information from the noisy data, thus improving classification accuracy.

We would also expect collective classification to perform better in networks that exhibit higher levels of homophily. Figure 2(a) shows how classification accuracy varies with different values of homophily (*dh* ranging from 0.4 to 0.9 with step size of 0.1). The results match our predictions: when homophily is low ($dh = 0.4$), the relational information only improves classification results slightly; but as we increase homophily, collective classification accuracy climbs steadily.

6.3 Aggregation Operators

Aggregation operators summarize the visible trust levels in a given agent's network neighborhood. In this set of experiments we explore the degree to which classification accuracy is affected by the choice of operator. We consider the following operators, each detailed below: *count*, *proportion*, *mode* and *reputation*.



(a) Changing link density with parameter setting *dh*=0.8,*attrNoise*=0.2

(b) Changing homophily with parameter setting *ld*=0.4,*attrNoise*=0.2

Fig. 2. The effect of changing link density (a) and homophily (b) on collective classification accuracy

As described earlier, *count* aggregates trust level labels of neighbors into a histogram of raw counts. *Proportion* is a normalized version of the *count* histogram. *Mode* retains only the most popular trust level, ranging from 1–4. Finally, *reputation* (as given in Equation 1) summarizes the agent's neighborhood in a single scalar quantity and can also be employed as an aggregation operator.

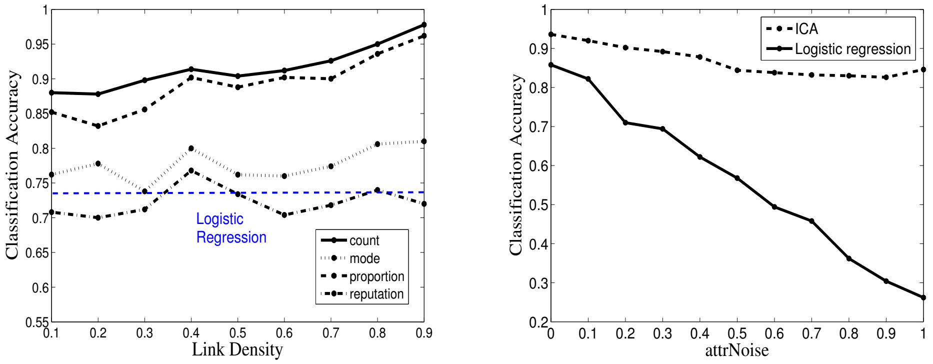
Figure 3(a) compares the classification accuracy of collective classification using the different aggregation operators against the LRC baseline. From the results, we make the following observations. First, compressing the relational information as a single scalar-valued *reputation* does not improve accuracy over the baseline. The *mode* operator is a little better, slightly but consistently outperforming the baseline. However, losing the richness of the visible trust levels (retaining only the most popular) is clearly inferior to the complete histogram of *proportion* or *count*. In fact, the unnormalized counts give the best results, and are therefore used as the default aggregation operator.

6.4 Robustness to Deception

So far, we have enforced a completely positive correlation between the agent's feature and its class label (trust level). However, in reality, cases may exist when certain untrustworthy individuals misrepresent themselves by modifying their observable information. In order to evaluate the performance of our model when this assumption is relaxed, we conduct two series of experiments. In the first experiment, we deliberately assign an increasing percentage of the deceptive nodes into the training dataset.

Here, the deceptive agent modifies its class label to appear more trustworthy (i.e., changing from L4 towards L1). Consequently, we select deceptive agents from classes *L2*, *L3*, and *L4*. We run 20 trials for each deception experiment with variable link density. Figure 4 shows the averaged results.

Collective classification (ICA) shows great robustness in this test (see Figure 4 and Table 4 . In a network with a modest amount of homophily, even when a large fraction of the population is deceptive (25% deceivers) ICA can continue to provide reliable results. It is important to note that employing collective classification on even a highly deceptive



(a) *dh*=0.8,*attrNoise*=0.2

(b) *dh*=0.8, *ld*=0.4

Fig. 3. Classification accuracy using (a) different aggregation operators; and (b) different noise values on synthetic trust dataset (*attrNoise*)

network is better than ignoring network information (ICA outperforms baseline of 75% in all conditions).

Table 4. Even with a high fraction of deceivers, using relations improves over the LRC baseline (75%). (*ld*=0.4,*dh*=0.8,*attrNoise*=0.2)

Deceivers (%)	1	5	10	15	20	25
Accuracy (%)	91.2	90.3	89.5	88.7	87.6	84.2

We also seek to explore the robustness of collective classification to a second form of deception: where the agent corrupts its observable features, generating a noisy observation vector. In our network generation model, the *attrNoise* parameter precisely captures the effect: each binary feature is randomized i.i.d. with a probability of *attrNoise*. As in earlier experiments, we compare collective classification (ICA) against the baseline (LRC), as shown in Figure 3(b). We make several observations. First, unlike in previous experiments, we confirm that the baseline accuracy decreases steadily as *attrNoise* rises, reaching chance level (25%) when $attrNoise = 1$. This is because an agent's observable features become an increasingly unreliable predictor of its trustworthiness. Second, by contrast we see that ICA's accuracy degrades surprisingly little, even when observable features become completely non-informative. This is because collective classification is still able to rely on network relations to predict an agent's trustworthiness based solely upon that of other agents in the neighborhood. Clearly, this can happen only when the network exhibits sufficient homophily and density.

7 Related Work

Trust evaluation has been applied to many diverse domains including peer-to-peer networks [18, 46], online social networks [48, 28, 34], e-business [29, 32] and mobile adhoc networks [4]. Identifying non trustworthy agents in multi-agent systems and coping

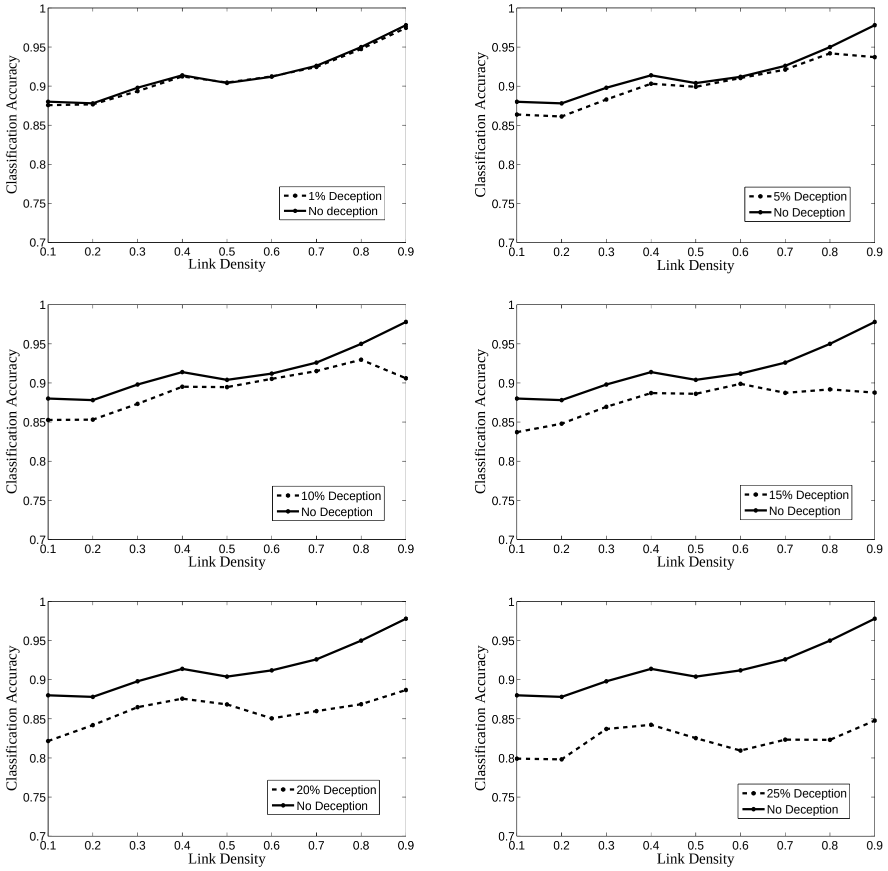


Fig. 4. Deception experiment using collective classification with the number of deceivers changing from 1% to 25%.

with the problem of cheating is important especially for the web and in electronic marketplaces. [7, 19] and [40] have proposed techniques to cope with cheaters and sneakers respectively. In our work, we are not only interested in identifying untrustworthy agents, but also finding highly trustworthy agents. Our approach uses local network information to perform a trust evaluation of other agents. In huge networks such as the Semantic Web, this local approach is also favored as the agents do not have access to all other agents. [48] offers some local metrics for trust and reputation in the Semantic Web domain.

Other authors have examined the relationship between trust and homophily in human social networks. Prisel and Anderson [33] observe that perceived homophily is positively related to feelings of safety and is negatively related to the level of uncertainty in groups. Evans and Wensley [9, 10] showed a direct link between homophily and trust; higher levels of status and value homophily increase the level of trust. They also note that homophily results in increased knowledge/information sharing activities across the group which are often a precursor to trust. However, status homophily has also been found to be negatively related to trust. In [21], the authors found no significant effect of status homophily on benevolence-based trust; age similarity was found to have a negative effect on competence-based trust. Overall, we believe that the link between trust and homophily is an interesting problem worthy of further study.

Our proposed trust evaluation approach identifies the correct label for all of the unlabeled agents in the network; this is the fundamental task of within-network classification techniques [8, 23]. Previous authors have looked at the problem of classifying nodes in social networks (e.g., [17, 16]). In these approaches, both network structure information and node class labels are combined to provide new features to improve classification [15]. Much of the previous work on using machine learning to identify the reputation or trust level of agents in a multi-agent system has used more traditional Bayesian methods (e.g., [12, 3]) and ignored the valuable information in the network structure information. We refer to the surveys of Macskassy et al. [23] and [2] for within-network classification techniques that have been used in social networks. Although, within-network classification has been used in fraud detection applications, such as call networks [11, 6], to detect the fraudulent or legitimate entities in the network, it has not been applied to problems of trust and reputation before. We believe that fraud-detection is another potential application for our trust evaluation approach. Our work is novel also in its detailed examination of the effects of agent deception on the classification performance of a collective classifier.

8 Conclusion

In this paper, we have demonstrated that when homophily in trustworthiness is a driving factor in the evolution of an agent network, collective classification is an effective mechanism for leveraging the informative powers of the network, even in the presence of other link generation forces such as transactional satisfaction. Although other types of supervised classifiers [44] and relational models of trust [37] have been explored, they do not propagate information across multiple instances to perform trust evaluation. Preserving the distribution of labels through more expressive aggregation operators such as count and proportion is shown to be more effective than the use of the single reputation feature that encodes the value differential between the trustworthiness of a node and its neighbors. In future work, we are particularly interested in applying this framework toward two types of problems: 1) using trustworthiness levels to perform link prediction in agent networks; 2) learning multi-dimensional models of trust from performance data.

Acknowledgments This research was supported in part by NSF award IIS-0845159.

References

1. A. Barabasi and E. Bonabeau. Scale-free networks. *Scientific American*, pages 60–69, May 2003.
2. S. Bhagat, G. Cormode, and S. Muthukrishnan. Node classification in social networks. *Arxiv preprint arXiv:1101.3291*, 2011.

3. J. Braams. Filtering out unfair ratings in Bayesian reputation systems. *The Icfain Journal of Management Research*, 4(2):48–64, 2005.
4. S. Buchegger and J. Le Boudec. A robust reputation system for mobile ad-hoc networks. *Proceedings of P2PEcon, June*, 2004.
5. C. Burnett, T. J. Norman, and K. Sycara. Bootstrapping trust evaluations through stereotypes. In *9th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 241–248, May 2010.
6. C. Cortes, D. Pregibon, and C. Volinsky. Communities of interest. *Intelligent Data Analysis*, 6(3):211–219, 2002.
7. C. Dellarocas. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49(10):1407–1424, 2003.
8. C. Desrosiers and G. Karypis. Within-network classification using local structure similarity. *Machine Learning and Knowledge Discovery in Databases*, pages 260–275, 2009.
9. M. Evans and A. Wensley. The influence of network structure on trust: Addressing the interconnectedness of network principles and trust in communities of practice. In *The 9th European Conference on Knowledge Management: ECKM 2008*, page 183. Academic Conferences Limited, 2008.
10. M. Evans and A. Wensley. Predicting the influence of network structure on trust in knowledge communities: Addressing the interconnectedness of four network principles and trust. *Electronic Journal of Knowledge Management*, 7(1):41–54, 2009.
11. T. Fawcett and F. Provost. Adaptive fraud detection. *Data mining and knowledge discovery*, 1(3):291–316, 1997.
12. R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Social network classification incorporating link type values. In *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on*, pages 19–24. IEEE, 2009.
13. T. Huynh, N. Jennings, and N. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.
14. M. Jackson. Social and economic networks, 2008.
15. T. Kajdanowicz, P. Kazienko, and P. Doskocz. Label-dependent feature extraction in social networks for node classification. *Social Informatics*, pages 89–102, 2010.
16. T. Kajdanowicz, P. Kazienko, and P. Doskocz. A method of label-dependent feature extraction in social networks. *Computational Collective Intelligence. Technologies and Applications*, pages 11–21, 2010.
17. T. Kajdanowicz, P. Kazienko, P. Doskocz, and K. Litwin. An sssessment of node classification accuracy in social networks using label-dependent feature extraction. *Knowledge Management, Information Systems, E-Learning, and Sustainability Research*, pages 125–130, 2010.
18. S. Kamvar, M. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*, pages 640–651. ACM, 2003.
19. R. Kerr and R. Cohen. Smart cheaters do prosper: Defeating trust and reputation systems. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 993–1000. International Foundation for Autonomous Agents and Multiagent Systems, 2009.
20. P. Lazarsfeld and R. Merton. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18:18–66, 1954.
21. D. Levin, R. Cross, and L. Abrams. Why should I trust you? Predictors of interpersonal trust in a knowledge transfer context. In *Academy of Management Meeting, Denver, CO*, 2002.
22. Q. Lu and L. Getoor. Link-based classification. In *In proceedings of 20th International Conference on Machine Learning*, pages 496–503. Association for Computing Machinery, August 2003.

23. S. Macskassy and F. Provost. A brief survey of machine learning methods for classification in networked data and an application to suspicion scoring. In *Proceedings of the 2006 conference on Statistical network analysis*, pages 172–175. Springer-Verlag, 2006.
24. S. A. Macskassy and F. Provost. A simple relational classifier. In *Proceedings of the Second Workshop on Multi-Relational Data Mining (MRDM-2003) at KDD-2003*, pages 64–76, August 2003.
25. S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. In *Journal of Machine Learning Research*, pages 935–983. Association for Computing Machinery, January 2007.
26. L. K. Mcdowell, K. M. Gupta, and D. W. Aha. Cautious inference in collective classification. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 596–601. AAAI Press, July 2007.
27. M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27:415–444, 2001.
28. L. Mui. *Computational models of trust and reputation: Agents, evolutionary games, and social networks*. PhD thesis, Massachusetts Institute of Technology, 2002.
29. L. Mui, M. Mohtashemi, and A. Halberstadt. A computational model of trust and reputation. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*, pages 2431–2439. IEEE, 2002.
30. J. Neville and D. Jensen. Iterative classification in relational data. In *Proceedings of the AAAI 2000 Workshop Learning Statistical Models from Relational Data*, pages 42–49, July 2000.
31. J. Neville and D. Jensen. Collective classification with relational dependency networks. In *Proceedings of KDD-2003 Workshop on Multi-Relational Data Mining (MRDM-2003)*, pages 77–91. AAAI Press, August 2003.
32. J. ODonovan, B. Smyth, V. Evrim, and D. McLeod. Extracting and visualizing trust relationships from online auction feedback comments. In *Proc. IJCAI'07*, pages 2826–2831, 2007.
33. M. Prisbell and J. Andersen. The importance of perceived homophily, level of uncertainty, feeling good, safety, and self-disclosure in interpersonal relationships. *Communication Quarterly*, 28(3):22–33, 1980.
34. J. Pujol, R. Sanguesa, and J. Delgado. Extracting reputation in multi agent systems by means of social network topology. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pages 467–474. ACM, 2002.
35. S. Ramchurn, D. Huynh, and N. Jennings. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19(01):1–25, 2004.
36. P. Resnick and R. Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *Advances in Applied Microeconomics: A Research Annual*, 11:127–157, 2002.
37. A. Rettinger, M. Nickles, and V. Tresp. A statistical relational model for trust learning. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, pages 763–770. International Foundation for Autonomous Agents and Multiagent Systems, 2008.
38. P. Sen and L. Getoor. *Link-based classification*. Technical Report, CS-TR-4858, University of Maryland, Reading, Massachusetts, 2007.
39. P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-rad. *Collective Classification in Network Data*. AI Magazine, 2008.
40. R. Seymour and G. Peterson. Responding to Sneaky Agents in Multi-agent Domains. In *Proceedings of the Florida AI Research Society Conference (FLAIRS)*, 2009.

41. B. Taskar. Discriminative probabilistic models for relational data. In *In Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence*, pages 485–492. Association for Uncertainty in Artificial Intelligence, August 2002.
42. G. van de Bunt, R. Wittek, and M. de Klepper. The evolution of intra-organizational trust networks. *International sociology*, 20(3):339–369, 2005.
43. Y. Wang and M. P. Singh. Formal trust model for multiagent systems. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1551–1556, January 2007.
44. W. Yuan, D. Guan, Y. Lee, and S. Lee. A trust model with dynamic decision making for ubiquitous environments. In *14th IEEE International Conference on Networks*, volume 1, pages 1–6. IEEE, 2007.
45. G. Zacharia and P. Maes. Trust management through reputation mechanisms. *Applied Artificial Intelligence*, 14(9):881–907, 2000.
46. R. Zhou, K. Hwang, and M. Cai. Gossiptrust for fast reputation aggregation in peer-to-peer networks. *Knowledge and Data Engineering, IEEE Transactions on*, 20(9):1282–1295, 2008.
47. X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning*, 2003.
48. C. Ziegler and G. Lausen. Propagation models for trust and distrust in social networks. *Information Systems Frontiers*, 7(4):337–358, 2005.

Trust Based Efficiency for Cake Cutting Algorithms

Roie Zivan and Ella Segev,
Industrial Engineering and Management department,
Ben Gurion University of the Negev,
Beer-Sheva, Israel
{zivanr,ellasgv}@bgu.ac.il

Abstract. Fair division methods offer guarantees to agents of the proportional size or quality of their share in a division of a resource (cake). These guarantees come with a price. Standard fair division methods (or "cake cutting" algorithms) do not find efficient allocations (not Pareto optimal). The lack of efficiency of these methods makes them less attractive for solving multi-agent resource and task allocation.

Trust can be the foundation on which agents exchange information and enable the exploration of allocations that are beneficial for both sides. On the other hand, the willingness of agents to put themselves in a vulnerable position due to their trust in others, results in the loss of the fairness guarantees.

In this work we extend the study on fair and efficient cake cutting algorithms by proposing a new notion of *trust-based efficiency*, which formulates a relation between the level of trust between agents and the efficiency of the allocation. In addition, we propose a method for finding trust-based efficiency. The proposed method offers a balance between the guarantees that fair division methods offer to agents and the efficiency that can be achieved by exposing themselves to the actions of other agents. When the level of trust is the highest, the allocation produced by the method is globally optimal (social welfare). [1]

1 Introduction

One of the main challenges in multi-agent systems (MAS) is encouraging self-interested agents to cooperate. Fair division methods offer a possible solution to this challenge for resource and task allocation, by offering guarantees to agents of the quality or size of their share, as long as they are cooperative (follow the instructions of the method's protocol). Moreover, these guarantees hold for an agent, even if other agents choose an uncooperative strategy.

The classic problem that is considered in fair division studies is the division of a heterogeneous resource (a cake) for which agents have their private utility/preference function. Agents divide the cake among themselves by performing *cut and choose* operations. The most familiar cut and choose method is dividing a cake between two agents, so that each will consider her share as at least half of the cake (a proportional share). The method requires one of the agents to cut the cake into two pieces which she considers

[1] A preliminary version of this paper was accepted for publication as an extended abstract of two pages in the proceedings of AAMAS 2011. This submission followed a query to the program chair of IJCAI 2011 who gave his approval.

to be equal and the other to choose the piece she prefers. It is obvious that both agents would consider their share to be at least half of the value of the entire cake. However, this method also demonstrates the weakness of fair division methods. The resulting allocation is guaranteed to be fair but might not be efficient (not Pareto optimal). In other words, there can exist a different allocation that is preferred by both (or preferred by one and is equal in the eyes of the other). For example, consider a round cake, half chocolate and half vanilla, and one agent who strictly prefers vanilla while the other strictly prefers chocolate. The agent cutting the cake may cut the cake so that each piece would include an equal amount of vanilla and chocolate. However, both agents would benefit from an allocation in which each agent gets her preferred flavor.

Many applications of resource and task allocation among self-interested agents motivate the study of methods for fair and efficient allocations. Task allocation in an industrial environment is one example where both fairness and efficiency are required. If, in the name of fairness, we allow workers to perform tasks that they are less qualified for than other workers, we lose efficiency and the resulting revenue of the factory is smaller. Such applications motivate the study of methods, which besides fairness guarantees, would offer guarantees on the level of efficiency.

Previous attempts to introduce efficiency into a fair division method offered extensions of *Austin's method* [2, 5]. Austin's method is the only method for finding an *exact* allocation of a cake among two agents, i.e., it finds an allocation where both agents consider the two pieces to be exactly half of the cake [12]. A simple extension to Austin's procedure increases the efficiency of the allocation in an asymmetric manner. One of the agents selects the most beneficial piece for herself such that the other agent considers her share as exactly half of the cake. This method has the following obvious limitations: (1) Only allocations that include up to two cuts of the cake are considered. (2) The method does not consider allocations in which both agents value their share as strictly more than 50%. A method that achieves a similar asymmetric increase in efficiency by allowing one agent to exploit a model she holds of the other agent's preferences, was proposed in [13]. This method has the same limitations as the asymmetric extension of Austin's method with the addition of the dependency on the existence of an accurate model held by one agent of the other agent's preferences.

The possibility of finding solutions to negotiation problems that *expand the pie*, i.e., the sum of the benefit for the negotiating parties exceeds 100%, was acknowledged by social scientists [15]. This acknowledgment triggered studies that investigated the success of different strategies in producing such agreements.

Trust is a concept that has been intensively studied by social scientists and by the multi-agent systems community [10]. The common and accepted definition of trust is the willingness of an agent to put herself in a situation in which she is vulnerable to the actions of another (the party she *trusts*). The relation between trust and efficiency was also acknowledged by multi-agent system studies [10]. In a cake cutting algorithm, it is easy to see how trust can increase the efficiency of the allocation. If the agents would exchange information regarding their preferences, they can reach an agreement in which each agent is allocated the parts she values more. On the other hand, sharing such information can put an agent in a vulnerable position. The other agents can exploit this knowledge to increase their own benefit. Thus, trust can allow agents to find efficient allocations, but at the same time, expose agents to the exploitive actions of others.

When examining realistic applications, situations in which there exists complete trust among self-interested agents are hard to find. It is common, for example, that people in a working environment would trust each other when they are working together on a project and share their ideas. However, rarely would an employee share her bank account details with her peers. Our approach towards trust is that there exists a scale on which the level of trust between agents can be measured and that the efficiency that can be reached by a cake cutting procedure can be incremented according to this level of trust. Notice, an agent may trust others to some extent to do the right thing in terms of global efficiency, even if it may result in her own loss. In other words, the agents do not trust one another to be fair but rather to be efficient.

In this paper we extend the research on fair and efficient cake cutting methods by:

1. Proposing a new notion of *trust-based efficiency*. It is a generalization of the concept of Pareto optimality, which reflects the level of trust between agents.
2. Proposing a method for finding *trust-based efficiency*. The method proposed allows agents to expose themselves with respect to the level of trust and make use of this exposure to increase efficiency while maintaining the guarantees on the fraction of the proportional share that the agents were not willing to risk. When the level of trust is maximal, the allocation found by the method is globally optimal (social welfare) [2].

Previous attempts to combine fairness and efficiency of a general cake considered the division of a cake between two agents (e.g., [13, 8]). This effort follows most studies on fair division, which attempt to solve the challenges such as proportionality, envy freeness, exact division, first for two agents and later propose a generalization to the case of n agents if possible [5, 12]. We follow this trend by formalizing the problem for the general case of n agents, proposing a solution for two agents and discussing the challenges that a generalized method will need to overcome.

2 Related work

Fair division is a well studied field that has drawn the attention of researchers for more than half a century [14, 5, 12]. The general aim of this field is to propose methods that allow agents with private preferences to divide a good among them. The methods offer guarantees to agents on the level of fairness, as long as they follow the protocol of the method. Standard fair division studies consider cake cutting algorithms in which agents perform the cut and choose operations to divide a heterogeneous resource (cake) among them [12].

Several studies acknowledged the existence of allocations that are both fair and efficient. Weller [16], and later Barbanel [3], prove the existence of envy-free Pareto

[2] Approximately fair protocols were suggested in previous studies, e.g., RobertsonW98. However, as far as we know, we are the first to present the relation between the level of guaranteed fairness and efficiency.

optimal allocations for a single heterogeneous good (cake) among n agents. However, there are very few studies on methods for finding fair and efficient allocations for a general cake. For multiple homogeneous divisible goods for which agents have linear utility functions, Reijnerse and Potters [11] propose an algorithm for finding an allocation among n agents, which is envy-free and Pareto optimal. Their solution is centralized, i.e., they assume that a central entity holds the true utility functions of all agents, and based on *market clearing* that is achieved using Fisher's model [7]. A later study [6] proposes a polynomial-time combinatorial algorithm for solving the same market clearing problem. The market clearing problem can also be solved using the *Eisenberg Gale* linear program [7].

Another attempt to apply a centralized algorithm for finding an efficient and fair division of multiple divisible homogeneous goods among two agents, was presented by Brams and Taylor [5]. They propose two methods that find an equitable allocation (an alternative notion of fairness in which both agents consider their allocation to be of the same value). One of the methods (the "adjusted winner") finds a Pareto optimal allocation while the other divides each of the goods proportionally between the two agents.

Another study, [8], proves that a division between two agents, which is fair and efficient with respect to a single planar cut of the cake, exists and offers a centralized method for finding it.

All of the above studies assume that a mediator holds the agents' preferences and computes the allocation. This is in contrast to standard cake cutting methods in which agents do not reveal their preferences to others [12].

The special case of an allocation of multiple indivisible goods has also drawn the attention of researchers. In this case a fair and efficient allocation does not always exist and thus, studies investigate the conditions for its existence and the complexity for finding it [9, 4].

We describe in Section 4 the method proposed by Sen and Biswas for increasing the efficiency of a division of a general cake between two agents via a cake cutting algorithm.

Trust is a concept whose different aspects are well studied by social scientists. These aspects include the development of trust [17] and the efficiency of teams with respect to the level of trust between their members [1]. The importance of the concept of trust in multi-agent systems was also acknowledged and drew extensive attention [10]. However, to the best of our knowledge, ours is the first attempt to increase the efficiency of resource allocation in multi-agent systems with respect to the level of trust, i.e., expose the agents partially with respect to the level of trust and use this exposure to increase efficiency.

3 Preliminaries

Our goal is to divide an infinitely divisible but bounded heterogeneous resource (cake) X between n agents. We assume that the cake has a rectangular shape with length L and width 1. We further assume that all cuts are planar. A piece of the cake x can be noted by an ordered pair $x = \langle x^l, x^r \rangle$, where $0 \leq x^l \leq x^r \leq L$. The numerical values of

x^l and x^r are their distances from the left edge of the cake (coordinates) and therefore the length of piece x is equal to $x^r - x^l$. Thus, $X = \langle 0, L\rangle$ and a result of a single cut at distance $c < L$ from the left edge of the cake is two pieces $\langle 0, c\rangle$ and $\langle c, L\rangle$. When $x^l = x^r$ the piece is empty (of size zero). For the operators $\subseteq$ and $\subset$ the standard definitions (for sets) apply for pieces as well.

The following operators are defined on pieces:

- $\tilde{\cap}$: assume without loss of generality that $m' \leq m$.
$$\langle m', n'\rangle \,\tilde{\cap}\, \langle m, n\rangle = \begin{cases} nil, & n' \leq m \\ \langle m, n'\rangle, & m < n' < n \\ \langle m, n\rangle, & n \leq n' \end{cases}$$
- $\tilde{\cup}$: assume without loss of generality that $m' < m$.
$$\langle m', n'\rangle \,\tilde{\cup}\, \langle m, n\rangle = \begin{cases} nil, & n' < m \\ \langle m', n\rangle, & m \leq n' < n \\ \langle m', n'\rangle, & n \leq n' \end{cases}$$

We will use the term sub-piece to describe a piece that is contained in another piece, i.e., $\langle m, n\rangle$ is a sub-piece of piece $\langle m', n'\rangle$ if and only if $\langle m, n\rangle \,\tilde{\subseteq}\, \langle m', n'\rangle$.

The operator $\tilde{\setminus}$ removes a sub-piece $\hat{x}$ from a piece x that contains it. The result is a set including the remaining two pieces to the left and right of the removed sub-piece. Formally:
$$\langle m, n\rangle \,\tilde{\setminus}\, \langle m', n'\rangle = \begin{cases} \{\langle m, m'\rangle, \langle n', n\rangle\}, & \langle m', n'\rangle \,\tilde{\subseteq}\, \langle m, n\rangle \\ \langle m, n\rangle, & \text{otherwise} \end{cases}$$

We define a *max-piece* in a set of pieces S as follows: $x \in S$ is a max-piece in S if there is no ordered subset $\{x_i, ..., x_k\}$ of S, for which $x \subset [x_i \tilde{\cup} ... \tilde{\cup} x_k]$. In other words, max pieces are obtained from a given set of pieces by applying the union operator on any two pieces that are not disjoint. Any set of pieces can be uniquely represented by a set of max pieces.

An allocation A is constructed of n disjoint finite sets of pieces, $X_1, ..., X_n$, such that if we order the pieces in the union of these sets according to their left coordinate (x^l) and apply the $\tilde{\cup}$ operator on all of them in this order (from left to right), we get the entire cake X. Furthermore, for any two pieces in this union, x and x', $x \,\tilde{\cap}\, x' = nil$. Intuitively, the entire cake is split between the agents and the cutting process does not decrease the quantities so the union of the agents' pieces is the entire cake.

We define a *max-allocation* to be an allocation in which all the pieces included in the sets X_a and X_b are max-pieces. Each allocation can be represented as a max allocation and this representation is unique. In the rest of this paper, when we discuss allocations, we will always refer to max-allocations unless we specifically say differently. Similarly, we will always refer to max-pieces when discussing pieces allocated to agents unless we specifically state differently

We further assume that for each agent i, $1 \leq i \leq n$ the function $F_i : \mathbb{R} \rightarrow \mathbb{R}$ defines for each point of the cake its value to agent i. We define $F_i(z) = 0$ for $z < 0$ and $z \geq L$. We assume that for $0 \leq z < L$, $F_i(z) > 0$ and that for $0 < z < L$, $F_i(z)$ is continuous and differentiable.

The utility function U_i defines the utility that agent i derives from a piece allocated to her, i.e., for $1 \leq i \leq n$:
$U_i(x) = \int_{x^l}^{x^r} F_i(z)\, dz$.

We assume that the utilities agents derive from an allocation of the entire cake are equal and we normalize them as follows:
$U_i(X) = \int_0^L F_i(z)\, dz = 1$.

We will use the notation $U_i(A)$ for the utility agent i derives from an allocation A, which is equal to the utility the agent derives from her allocated set of pieces in A,
$U_i(A) = \sum_{x \in X_i} \int_{x^l}^{x^r} F_i(z)\, dz$.

We assume that an agent i can compute accurately for any $0 \leq m \leq n \leq L$ the integral: $\int_m^n F_i(z)\, dz$ and that agents can perform cuts accurately, i.e., if an agent cuts a piece $\langle m, n \rangle$ at point k, $m \leq k \leq n$, the result are two pieces $\langle m, k \rangle$ and $\langle k, n \rangle$.

We will call an allocation A *efficient*, if it is Pareto optimal, i.e., there is no other allocation A' so that for some $j \in \{1, ..., n\}$: $U_j(A') > U_j(A)$ and for all $1 \leq i \leq n, i \neq j$: $U_i(A') \geq U_i(A)$.

The *social welfare value* of an allocation A is the summation of utilities $U_1(A) + ... + U_n(A)$. An allocation A has an optimal social welfare value (SW_{opt}) when there is no A' with $U_1(A') + ... + U_n(A') > U_1(A) + ... + U_n(A)$.

An allocation A is *proportional* if for every agent $1 \leq i \leq n$, $U_i(A) \geq \frac{1}{n}$.

4 Austin's method and asymmetric extensions

Austin's moving knife procedure can find an *exact* division of a heterogeneous cake between two agents, in which both agents consider their share as exactly $\frac{1}{2}$ [2, 12]. One agent (without loss of generality we will assume that this is agent a) holds two parallel knives. In the initial state, the left knife is placed at point zero and the right knife at point r so that $\int_0^r F_a(z)\, dz = \frac{1}{2}$. Agent a moves both knives to the right so that for every location of the left knife, the right knife is placed so the piece between the knives is worth exactly $\frac{1}{2}$ to her (we will refer to the piece between the knives as P and to the remainder of the cake as P'. P_{ll} and $\bar{P_{ll}}$ will be used to note the piece between the knives and the remainder when the left knife location is ll. The initial location of the left knife is 0, in which agent a puts the left knife on the left edge of the cake. The final location in which the right knife reaches the right edge will be noted by $\hat{ll}$. When at some location of the left knife ll, $0 \leq ll \leq \hat{ll}$, $U_b(P_{ll}) = \frac{1}{2}$, agent b calls stop, she gets P_{ll} while agent a gets $\bar{P_{ll}}$. Notice that if both agents followed the protocol, $U_a(P_{ll}) = U_b(\bar{P_{ll}}) = \frac{1}{2}$. Such an allocation can always be found by Austin's procedure due to the continuous nature of the scan of the two knives by agent a.

A small adjustment to Austin's procedure can result in increased efficiency. Notice that while $U_a(P_{ll}) = \frac{1}{2}$ for any location of the left knife ll, $0 \leq ll \leq \hat{ll}$, $U_b(P_{ll})$ may be changing. Thus, if we would allow agent b to observe the full process in which agent a moves the knives from the initial position to the final complementary position, and then choose the piece $P_{ll}, 0 \leq ll \leq \hat{ll}$, we can increase the efficiency of the method, since for the resulting allocation A', $U_b(A') \geq \frac{1}{2}$ while $U_a(A') = \frac{1}{2}$. However, it is clear that this increment in efficiency is one-sided (agent a would never derive more utility than $\frac{1}{2}$).

A different extension to Austin's method, which increases its efficiency, was proposed by Sen and Biswas [13]. This method is also asymmetric, only in contrast to the asymmetric extension of Austin's method described above, here, the advantage is to the cutting agent (a). The advantage for agent a is derived from the assumption that she is holding a model of the utility function of agent b, $\hat{U}_b$. As before, we assume that agent b is allowed to observe the entire process in which the knives are moved by agent a across the cake and select the position of the left knife ll, in which $U_b(P_{ll})$ is maximal.

In order to increase the efficiency of the allocation, agent a selects a piece $\hat{P}$, for which $\hat{U}_b(\hat{P}) = \frac{1}{2}$ and $U_a(\hat{P})$ is minimal. Then, she makes sure that agent b will prefer this piece $\hat{P}$ over any other P_{ll} by keeping the knives so that $\hat{U}_b(P_{ll'}) < \frac{1}{2}$ for $ll' \neq ll$. Thus, agent b selects $\hat{P}$ and if $\hat{U}_b$ is accurate, the utility for agent a is the greatest possible among the allocations with two cuts in which agent b receives a proportional share.

The two methods described above are both asymmetric. Both give an advantage to one of the agents over the other. This advantage allows the agent to choose the allocation that maximizes her gain, given that the allocation does not require more than two cuts and that the utility the other agent derives from it is exactly $\frac{1}{2}$. However, the utility derived from the allocation to the agent who does not have the advantage can never be larger than $\frac{1}{2}$. Therefore, allocations that increment the benefit for both agents are not considered.

5 Trust Efficient Allocations

The shortcoming of asymmetric methods in finding allocations that extend the benefit for both agents beyond their proportional share, motivates the development of a model or method that will enable such allocations. As mentioned in Section 1, the relation between trust and efficiency in applications such as multi-agent negotiation, has been acknowledged in previous studies. However, besides the potential for cooperation between agents that will result in efficiency, by definition, trust includes the willingness of agents to become vulnerable to the manipulations of other agents. Such vulnerability somewhat contradicts the motivation for fair division methods that offer guarantees to agents regardless of the actions of others. In reality, this trust is rarely a "take it or leave it" (binary) choice. While it would not be realistic to assume that an agent would trust another enough to risk her entire share, commonly, some level of trust between agents does exist. In other words, in many cases agents would be willing to expose themselves partially in order to increase the efficiency of the result. The amount of risk they will be willing to take (the level of trust) is determined by many elements and has been studied by social scientists [17]. Our goal is to introduce trust into the existing efficiency formalization of cake allocations. This formalism will set lower bounds on the efficiency of allocations dependent on the level of trust between agents.

We propose a novel approach to efficiency in cake cutting algorithms depending on the level of trust between the agents. To this end, we make the following innovative definitions:

Definition 1 *l-trust: given l, the symmetric level of trust among agents $1, ..., n$, l-trust is the fraction of the proportional share that the agents are willing to risk.*

Following this definition is an incentive participation constraint for each agent $1 \leq j \leq n$, where for any possible resulting allocation A, $U_j(A) \geq \frac{1-l}{n}$.

Definition 2 *l-trust-efficiency: An allocation A is l trust efficient if there does not exist an ordered set of agents $1, ..., k$ each holding max-pieces $x_j, 1 \leq j \leq k$, respectively, such that $U_j(x_{j-1}) \geq U_j(x_j)$ for $j = i+1, ..., k$ and $U_i(x_k) > U_i(x_i)$. Furthermore, $U_j(x_{j-1}) \geq \frac{1-l}{n}$, $j = i+1, ..., k$ and $U_i(x_k) \geq \frac{1-l}{n}$.*

Intuitively an l-trust-efficient allocation does not include a *Pareto improvement exchange cycle*, which is a cycle in which each of the participating agents gives a piece to another agent and receives a different piece, and for one agent this exchange increases her utility while for all others the utility does not decrease [3]. For l-trust-efficiency we add another constraint, that the derived utility for each agent from the piece she is receiving is at least $\frac{1-l}{n}$,

The definition of l-trust-efficient allocations is inspired by the definition of Pareto optimal allocations in which no exchange that is strictly beneficial to one agent and weakly beneficial to all others is possible [3]. Intuitively, l-trust-efficiency is the resolution in which the value of pieces to different agents can be identified.

6 Finding l-trust-efficient allocations between two agents

For two players we start by extending the problem definition from Section 3 to include l-trust. Formally, we assume that besides the cake X and the functions F_a and F_b, the input of the problem includes the symmetric level of trust, l, $0 \leq l < 1$. Our aim is to find an l-trust-efficient allocation A, in which $U_a(A) \geq \frac{1-l}{2}$ and $U_b(A) \geq \frac{1-l}{2}$.

6.1 LTE

We propose the following method for finding l-trust-efficient allocations, LTE:

1. Agent a places the left knife on the left edge of the cake and the right knife so that $U_a(P_0) = \frac{1-l}{2}$.
2. Agent a moves the knives to the right, keeping $U_a(P_{ll}) = \frac{1-l}{2}$ until the right knife reaches the right edge of the cake.
3. Agent b decides which pieces of the cake to allocate to agent a and which parts to herself, cuts the cake and makes the allocation accordingly.

It remains to describe how agent b decides on the allocation at the third step. Notice that, like in Austin's procedure, while $U_a(P_{ll})$ remains the same for each ll, $0 \leq ll \leq \hat{ll}$, $U_b(P_{ll})$ may be changing with the movement of the knives. The function $U_b(P_{ll})$ is observed and analyzed by agent b, in order to produce the allocation.

If possible, agent b selects a value $v \leq \frac{1-l}{2}$ and selects a set of pieces X_a where:

1. $x \in X_a \Rightarrow x = P_{ll}, 0 \leq ll \leq \hat{ll}$.
2. $\forall x, x' \in X_a, x \neq x' \Rightarrow x \tilde{\cap} x' = nil$.
3. $x \in X_a \Rightarrow U_b(x) \leq v$.

4. $\exists \hat{x} \subseteq x \in X_b$ s.t. $\hat{x} = P_{ll}, 0 \leq ll \leq \hat{ll} \Rightarrow U_b(\hat{x}) > v$.
5. $U_b(X_b) \geq \frac{1-l}{2}$.
6. $X_a \neq \emptyset$.

Notice that the conditions above do not necessarily define a max-allocation since the pieces in X_a can be adjacent. However, as always the resulting allocation has an equivalent max-allocation.

If no such value v can be found, then agent b selects any location ll, $0 \leq ll \leq \hat{ll}$ for which $U_b(P_{ll})$ is minimal and allocates P_{ll} to a, leaving the rest of the cake for herself.

The conditions listed above for selecting the value v offer some degree of freedom for agent b in selecting an l-trust-efficient allocation. We will call the method in which the maximal possible value for v is selected *LTE-max*. We will prove in Section 6.2 that the selection of the maximal value for v maximizes the social welfare value of the resulting allocation. We note that selection of a value v can result in a number of possible allocations. In order to establish determinism we further assume that the following two ordering decisions are used when selecting the pieces that will be added to X_a:

1. A piece x will always be added to X_a before a piece x' if $U_b(x) < U_b(x')$.
2. If $U_b(x) = U_b(x')$, then the piece with the smaller left coordinate will be selected first.

When there is a limit to the number of cuts that can be made when performing the allocation, the LTE method can be adjusted by adding an additional constraint to the conditions for the selection of value v. The cuts should be made to generate max-pieces only after the max-allocation is identified. Thus, assigning a set of consecutive pieces to a single agent would result in two cuts at most. The smallest number of cuts that allows LTE to find an l-trust-efficient allocation is 2. These two cuts are required so that at least one piece x with $U_a(x) = \frac{1-l}{2}$ and with minimal value b, can be allocated to agent a.

6.2 Properties

The first property is concerned with the guarantees provided to agents by the LTE method.

Theorem 1 *For any allocation A found by LTE,* $U_a(A) \geq \frac{1-l}{2}$ *and* $U_b(A) \geq \frac{1-l}{2}$.

Proof: Immediate by construction.

Next, we prove that the LTE method proposed above indeed finds an l-trust-efficient allocation.

Theorem 2 *Any allocation found by LTE is l-trust-efficient.*

Proof: The case where no value v that satisfies the conditions described in Section 6.1 exists is trivial; therefore we will only prove the case in which such a value v was found.

Assume that there exists a piece $x \in X_a$ and a piece $x' \in X_b$ so that (reminder, we are considering max-allocations):

1. $U_a(x') \geq \frac{1-l}{2}$.
2. $U_b(x) \geq \frac{1-l}{2}$.
3. $U_a(x) < U_a(x')$.
4. $U_b(x) \geq U_b(x')$.

By construction, $U_a(x) = k\frac{1-l}{2}, k \in \mathbb{N}$, and $U_b(x) < kv$. Therefore, according to the assumption, $U_a(x') > k\frac{1-l}{2}$ and thus, x' can be divided into $k+1$ consecutive sub-pieces, where for each sub-piece $\hat{x}'$ among the first k, $U_a(\hat{x}') = \frac{1-l}{2}$ and $U_b(\hat{x}') \geq v$. Thus, $U_b(x') \geq kv > U_b(x)$ in contrast to our assumption.

Notice that in the last expression we do not use the additional $k+1$ sub-piece. Therefore, the same proof holds for the case where:

1. $U_a(x') \geq \frac{1-l}{2}$.
2. $U_b(x) \geq \frac{1-l}{2}$.
3. $U_a(x) \leq U_a(x')$.
4. $U_b(x) > U_b(x')$.

□

The selection of v can affect the social welfare value. Therefore, we prove the following property:

Theorem 3 *For two allocations A and A' found by LTE with the corresponding values v and v', $v \geq v' \Rightarrow U_a(A) + U_b(A) \geq U_a(A') + U_b(A')$.*

Proof: Since $v \geq v'$, either $A = A'$, or there exists a piece (without loss of generality we assume there is exactly one such piece) x with the following properties:

1. $U_a(x) = \frac{1-l}{2}$.
2. $U_b(x) \leq v \leq \frac{1-l}{2}$.
3. in A, $x \in X_a$.
4. in A', $x \in X_b$.

Therefore, due to the deterministic manner in which the allocation to agent a by agent b is determined in LTE, both allocations are identical for $X\tilde{\setminus}x$ and $U_a(x) > U_b(x)$. Thus, $U_a(A) + U_b(A) \geq U_a(A') + U_b(A')$. □

Last, we prove the strong relation between our proposed notations and method to global efficiency (social welfare value). We start by defining a *flip point*: $\hat{z}$ is a flip point if $F_j(\hat{z}) = F_i(\hat{z})$ and $F'_j(\hat{z}) \neq F'_i(\hat{z})$.

We state the following Lemmas:

Lemma 1 *In LTE, a piece x for which $U_a(x) \geq \frac{1-l}{2}$ or $U_b(x) \geq \frac{1-l}{2}$, and x does not contain a flip point, is allocated to agent a if and only if $U_a(x) \geq U_b(x)$.*

Lemma 2 *For a piece of the cake $\langle z^l, z^r \rangle$, which does not contain a flip point, and $F_a(z) > F_b(z)$ for each $z^l \leq z \leq z^r$, the number of sub-pieces that are not allocated to agent a is equal to the number of extreme points of F_b in $\langle z^l, z^r \rangle$.*

Now we can state and prove the following theorem:

Theorem 4 *When the level of trust between agents is the highest, LTE-max finds an allocation that is optimal in terms of social welfare, i.e.,* $[U_a(A_l) + U_b(A_l)]_{l \to 1} = SW_{opt}$ [3].

Proof: Note by A^* the allocation that maximizes social welfare. Assume there are k flip points in X. According to Lemma 1, if piece x does not contain a flip point and one of the agents values it at least as $\frac{1-l}{2}$, then LTE-max allocates x to the agent who values it more. Thus, all P_{ll} pieces, $0 \leq ll \leq \hat{ll}$, in A (an allocation found by LTE-max) that do not contain flip points are allocated as in A^* except for the k pieces, which include the flip points and the pieces in set M from which agent a derives less utility than $\frac{1-l}{2}$, which are allocated to agent b. According to Lemma 2 this number is bounded by the number of extreme points in F_b. Thus, the following holds: $U_a(A_l) + U_b(A_l) \geq U_a(A^*) + U_b(A^*) - k\frac{1-l}{2} - |M|\frac{1-l}{2} = U_a(A^*) + U_b(A^*) - (k + |M|)\frac{1-l}{2}$.
Since $lim_{l \to 1}\frac{1-l}{2} = 0$, we get that when $l \to 1, U_a(A_l) + U_b(A_l) \geq U_a(A^*) + U_b(A^*)$.
□

Notice that the scale of the level of trust has social welfare on one side (when $l \to 1$); on the other side, when $l = 0$ we get the asymmetric extension of Austin's method.

6.3 Example

Consider the example depicted in Figure 1. The agents have contradicting preferences for the left side of the cake, while having similar preferences for its right side. We evaluated two levels of trust, $l = 0.4$ and $l = 0.8$, and had agent a move the knives accordingly. The functions $U_b(P_{ll})$, which agent b generates for the two different levels of trust, are depicted in Figure 2. The utilities derived by the agents from the resulting allocations, are depicted in Table 1.

The sum of the resulting utilities that the agents derive from the allocation ($U_a = 0.3$ and $U_b = 0.8$) expands the benefit beyond 100% as derived in Austin's procedure. However, agent a received the minimal value according to the l-trust guarantees.

l-trust	0.4	0.8
U_a	0.3	0.7
U_b	0.8	0.719
SW	1.1	1.419

Table 1. Utilities derived by agents and the social welfare value in the different example scenarios

If the trust level between the agents is greater, e.g., $l = 0.8$, agent b can be much more specific and expressive regarding her preferences. The resulting utilities are $U_a(X_a) = 0.7$ and $U_b(X_b) = 0.719$. Thus, the greater level of trust not only enabled an allocation with greater social welfare value (1.419), but also, both agents derived utilities beyond a 50% allocation.

[3] A_l is defined as before, an allocation found by LTE when the level of trust is equal to l.

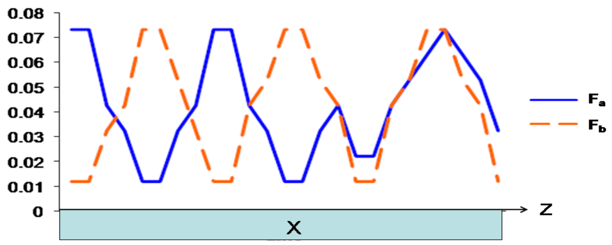


Fig. 1. Example of LTE

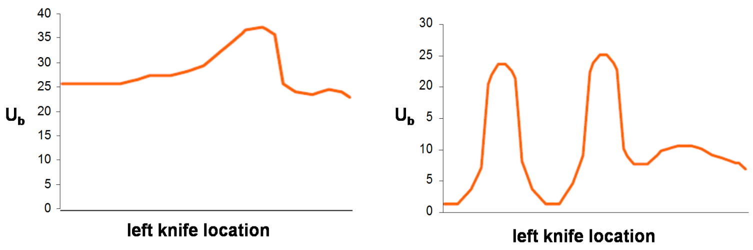


Fig. 2. $U_b(P_{ll})$ for $l = 0.4$ (left) and $l = 0.8$ (right)

7 Discussion of the general case

An l-trust-efficient allocation can be found using the following extension of the proposed LTE method for two agents. We order the agents $1, ..., n$ and let the first among them, agent 1, move the knives in the same manner as agent a did in the two agents method. Each agent $1 \leq i \leq n-1$ generates the function $U_i(ll)$ as done by agent b in the two agents version. Then, all functions $U_i(ll)$ are passed to agent n, which generates the function $U_{max}(ll) = maxU_i(ll), 1 \leq i \leq n-1$. Agent n adds to X_n disjoint pieces so that $x \in X_n \Rightarrow U_i(x_l) = U_{max}(x_l)$ and $U_n(x) > \frac{1-l}{n}$. The pieces are added in a deterministic order beginning with the piece from which agent n derives the most utility to the piece which she derives the least from. This process repeats with agent $n-1$ selecting pieces not yet allocated according to $U_{max}(ll)$, and so on until finally agent $n-1$ splits what ever is left with agent n.

While this method would result in an l-trust efficient allocation of pieces x for which $U_1(x) = \frac{1-l}{n}$, the following issues need to be solved so the generalized method will apply to the properties achieved by the two agent method:

1. There can exist a piece x', for which $U_1(x') < \frac{1-l}{n}$ but for some other agent $j > 1$, $U_j(x') = \frac{1-l}{n}$. We need to be careful when allocating such pieces in order not to lose l-trust-efficiency.
2. The allocation may not satisfy the trust guarantees for some agents, i.e., when agent i is considering her share, there might be not enough left so that $U_i(X_i) \geq \frac{1-l}{n}$. We

will need to propose some initial phase in which each agent receives her guarantee before applying the method above.

8 Conclusion

In this paper we proposed the use of *trust in cake cutting algorithms*. We defined the level of trust between agents as the proportional quantity of their fair share that they are willing to expose to the actions of other agents, and risk losing. We further defined a new concept, l-trust-efficiency, which generalizes the Pareto efficiency concept. When an allocation is l-trust-efficient, there does not exist any other allocation that can be derived from the current allocation by exchanging pieces that are worth at least $\frac{1-l}{n}$ to the agents between them and is strictly better for at least one agent and at least as beneficial to all other agents as the current allocation.

We proposed a method for finding l-trust-efficient allocations between two agents. The method allows agents to achieve this kind of efficiency with respect to the level of trust between them, but at the same time, guarantees the allocation of the quantity that they were not willing to risk. The method allows the agents to divide the cake between them according to the utility they derive from allocations of the different parts of the cake (the one who values it more gets the share) and, as a result, achieve not only the efficiency we defined but also increased the social welfare of the allocation.

We discussed the challenges in proposing a method that finds l-trust-efficient allocations between n agents. In future work we intend to find solutions to these challenges and propose a method for the general case.

References

1. B. E. Ashforth and R. T. Lee. Defensive behavior in organizations: A preliminary model. *Math. Gazett*, 43:621–648, 1990.
2. A. K. Austin. Sharing a cake. *Math. Gazett*, 66:212–215, 1982.
3. J. Barbanel. Partition ratios, pareto optimal cake division, and related notions. *Journal of Mathematical Economics*, 32:401–428, 1999.
4. S. Bouveret and J. Lang. Efficiency and envy-freeness in fair division of indivisible goods: Logical representation and complexity. *Journal of Artificial Intelligence Research (JAIR)*, 32:525–564, 2008.
5. S. J. Brams and A. D. Taylor. *Fair Division. From cake-cutting to dispute resolution*. Cambridge University press, 1996.
6. N. R. Devanur, C. H. Papadimitriou, A. Saberi, and V. V. Vazirani. Market equilibrium via a primal-dual-type algorithm. In *FOCS '02: Proceedings of the 43rd Symposium on Foundations of Computer Science*, pages 389–395, Washington, DC, USA, 2002.
7. D. Gale. *The Theory of Linear Economic Models*. McGraw-Hill, 1960.
8. M. A. Jones. Equitable, envy-free, and efficient cake cutting for two people and its application to divisible goods. *Mathematics Magazine*, 75(4):275–283, 2002.
9. M. Meertens, J. A. M. Potters, and J. H. Reijnierse. Envy-free and pareto efficient allocations in economies with indivisible goods and money. *Mathematical Social Sciences*, 44(3):223–233, 2003.
10. S. D. Ramchurn, D. Huynh, and N. R. Jennings. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19(1):1–25, 2004.

11. J. H. Reijnierse and J. A. M. Potters. On finding an envy-free pareto-optimal division. *Mathematical Programming*, 83:291–311, 1998.
12. Y. J. Robertson and W. Webb. *Cake-Cutting Algorithms, Be Fair If You Can*. A K Peters, Ltd, 1998.
13. S. Sen and A. Biswas. More than envy-free. In *ICMAS '00: Proceedings of the Fourth International Conference on MultiAgent Systems (ICMAS-2000)*, page 433, Washington, DC, USA, 2000. IEEE Computer Society.
14. H. Steinhaus. The problem of fair division. *Econometrica*, 16:101 – 104, 1948.
15. L. R. Weingart and J. M. Brett. Tactical behavior and negotiation outcomes. *International Journalof Conflict Management*, 1:7–31, 1990.
16. D. Weller. Fair division of a measurable space. *Journal of Mathematical Economics*, 14(1):5–17, February 1985.
17. J. M. Wilson, S. G. S. b, and B. McEvily. All in due time: The development of trust in computer-mediated and face-to-face teams. *Organizational Behavior and Human Decision Processes.*, 99(1):16–33, 2002.