A Situation-aware and Social Computational Trust Model

Universidade do Porto

Faculdade de Engenharia

FEUP

Maria Joana Malaquias Pires Urbano

Faculty of Engineering, University of Porto
Department of Informatics Engineering

July 2013

A Situation-aware and Social Computational Trust Model

Maria Joana Malaquias Pires Urbano

Supervisor: Professor Doutor Eugénio da Costa Oliveira
Co-Supervisor: Prof. Doutora Ana Paula Rocha

Thesis submitted in partial fulfillment of the requirements for the Doctoral degree in Informatics Engineering.

This work was supported by Fundação para a Ciência e a Tecnologia (FCT) through grant FCT SFRH/BD/39070/2007

To my parents

Acknowledgments

Life is made of decisions and deciding to leave my home town and start my PhD at the University of Porto was a wise one. Now it is the time to thank all people that helped to make this a fruitful and pleasant journey. First of all, I would like to thank my supervisors Professor Eugénio Oliveira and Professor Ana Paula Rocha for all the support I have received, for all insightful discussions and guidance; Henrique Lopes Cardoso, with who it was a pleasure to work with, and also, an inspiration; all team of the ANTE project and the people at our Lab room, António Castro, António Pereira and Pedro Brandão, with whom I had the pleasure to share lots of meals and good time, but also Célia, Gustavo, Leonardo, Luís, Sérgio, Jorge, Luís Paulo and Rosaldo. A special thanks to my cousins Filipa and Diana, who have also made me feel at home since the first moment.

However, I cannot forget my Coimbra's team: my parents, for their continuous and precious support since ever. My siblings and siblings-in-law, my dearest nieces, and all close family. Many thanks to the Direction of ISMT for their institutional support, and to all friends at ISMT, with a special thanks to Sofia. And a word of gratitude to my closest friends, with a special thanks to Jorge and Nuno for always being present. Finally, I would like to thank Gonçalo for (his sometimes nagging way of) encouraging me to do better each day.

The work presented in this document was supported by Fundação para a Ciência e a Tecnologia (FCT), under project PTDC/EIA-EIA/104420/2008 and PhD grant SFRH/BD/39070/2007.

Abstract

Computational trust modeling is a research field with a fast growing development since the last decade, mainly in the scientific area of distributed artificial intelligence and multi-agent systems, and its potential applicability spreads from social networks to distributed resource sharing and electronic markets. The earliest approaches to computational trust addressed the development of algorithms to aggregate the evidence on any given agent under evaluation into an estimated score of this agent's trustworthiness. More recently, the research on computational trust has shifted to the inclusion of third-party information about the trustee under evaluation, including opinions and reputation. However, one important aspect of computational trust has being neglected all these years by the majority of the scholars on computational trust, with a few relevant exceptions: trust is a social construct with a cognitive and an emotional account, and it strongly depends on the relationship existing between the agent that trusts and the one that is trusted.

In this thesis, we address the topic of social trust and its consideration for application in computational trust. We first present a thorough multidisciplinary view of trust, and derive important propositions that will guide our work throughout the thesis. Based on these propositions, we present the SOLUM model, our proposal to computational trust comprised of two distinct parts. The first part is a general framework of computational trust that is based on two fundamental characteristics of trust: trust is more than trustworthiness and other important antecedents to trust, such as the truster's disposition and emotional state, must be considered when estimation the truster's trust; and trustworthiness is a multi-dimensional construct that includes the ability, integrity, and benevolence dimensions. This framework can be instantiated and applied to a wide range of trust-based problems and applications, and is seen here as the first main contribution of this thesis.

The second part of the SOLUM model includes a set of distinct computational components that (partially) instantiate the framework, namely:

Sinalpha, *Contextual Fitness*, *Social Tuner*, and *Integrity Tuner*. We propose the use of specific techniques to extract information about the individual dimensions of the agents' trustworthiness from the set of structured evidence available on these agents, which may be scarce. This constitutes an innovative view over computational trust, and is the second main contribution of this thesis.

We evaluated our approach through experimental simulation. The results of our experiments allowed us to conclude that it is possible to improve in a relevant way the reliability of the trustworthiness estimations – for the same set of evidence – using the proposed techniques. Consequently, our approach contributes for more informative and secure decisions, in all domains where computer-based trust decisions are needed.

Resumo

A investigação em confiança computacional tem sofrido um rápido desenvolvimento ao longo da última década, principalmente nas áreas da inteligência artifical distribuída e dos sistemas multi-agente.

As primeiras abordagens à confiança computacional focaram-se no desenvolvimento de algoritmos capazes de agregar o conjunto de evidências existentes acerca de um determinado agente em avaliação de forma a calcular um valor estimado para a confiabilidade desse agente. Mais recentemente, a investigação nesta área evoluiu para a agregação de evidências utilizando informação de confiança obtida a terceiros, a partir, por exemplo, de opiniões e reputação. Pese embora todos estes avanços, há um aspecto ligado à confiança computacional que tem sido negligenciado pela maior parte dos académicos nesta área, com uma ou duas relevantes excepções: a confiança é um conceito social de construção cognitiva e social, que é fortemente dependente da relação existente entre o agente que confia e aquele que é objecto de confiança.

Nesta tese, nós endereçamos a confiança como conceito social e a forma como este conceito pode ser transportado para o domínio da confiança computacional. Assim, começamos por apresentar uma visão abrangente e multidisciplinar do conceito de confiança, visão esta que é depois capturada num conjunto de proposições que irão conduzir o nosso trabalho ao longo da tese. Tendo como base estas proposições, apresentamos o modelo SOLUM, que constitui a nossa proposta para um modelo de confiança computacional, e que é constituído por duas partes distintas.

A primeira parte é uma plataforma genérica para a confiança computacional, baseada em duas características chave da confiança: confiar é algo mais do que a previsão da confiabilidade do agente avaliado, pelo que avaliar a confiança de um agente noutro agente implica também, por exemplo, estimar a prepensão para confiar e o estado mental daquele que confia; da mesma forma, a confiabilidade do agente avaliado deve ser medida tendo em conta a competência desse agente na tarefa em causa, assim como a sua integri-

dade e a sua benevolência para com o agente avaliador. A plataforma assim desenvolvida pode ser instanciada e aplicada a um conjunto abrangente de problemas e aplicações que envolvam, de alguma forma, decisões baseadas em confiança, constituindo, deste modo, um dos contributos mais importantes desta tese.

A segunda parte do modelo SOLUM engloba um conjunto de componentes computacionais que instanciam, ainda que de forma parcial, a nossa plataforma. Estes componentes chamam-se *Sinalpha*, *Contextual Fitness*, *Social Tuner* e *Integrity Tuner*. No contexto do desenvolvimento destes componentes, propomos o uso de técnias específicas capazes de extrair informação sobre a competência, a integridade e a benevolência do agente avaliado a partir da informação existente acerca desse agente, que normalmente existe em pouca quantidade e encontra-se representada de forma estruturada. Esta é uma parte bastante inovadora da nossa tese, e constitui o segundo contributo importante da tese.

A nossa abordagem à confiança computacional foi avaliada por avaliação experimental. Os resultados das experiências efectuadas permitiu-nos concluir que é possível melhorar, de forma relevante, a fiabilidade do cálculo da confiabilidade dos agentes, para um mesmo conjunto de evidências, usando as técnicas que propomos. Consequentemente, o nosso trabalho aqui exposto dá um contributo importante para a tomada de decisões mais informadas e seguras em todos aqueles domínios onde são necessárias decisões baseadas em confiança.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the advent of the digital economy, relationships between business partners are increasing in flexibility, and business binds tend to be created whenever a business opportunity arises. Also, in order to cope with the emergent need for new products and services, with increased quality, short time to market, and low price, enterprises tend to try new, sometimes unknown, suppliers, possibly spread all over the world. This new reality brings new technological, social, ethical, and economical challenges and risks to enterprises.

In business-to-business (B2B) environments, even the most flexible approaches assume some degree of rigidness and technological commitment in the establishment of interactions between the business partners. In this respect, a branch of the distributed artificial intelligence (DAI) community has being developing models and infrastructures aiming at supporting the life-cycle of virtual organizations, i.e., the temporary aggregation of legal and independent organizations that share resources and competencies via a communication network in order to reach a given mission or global objective (Oliveira and Rocha, 2001; Camarinha-Matos et al., 2004). The paradigm of virtual organizations is growing in interest for business players as worldwide markets' instability claims for means to rapidly explore new opportunities and, at the same time, to easily extinguish the entire operational infrastructure associated with these opportunities when it is no longer needed. This process of creation and dissolution of partnerships is considered by some authors as more advantageous within the virtual organization framework than the traditional collaborative models, such as mergings, acquisitions and joint ventures (e.g., Dang, 2004). Also, it allows for a more flexible introduction of new partners, whenever there is the need to provide extra service. Some examples of potential use of virtual organizations include international oil

exploration consortia (Grabowski and Roberts, 1999), high-tech chip industry (Haller, 2008), and the construction/building industry (Karabulut and Sairamesh, 2005). More interesting, this possibility of rapidly configuring a collaborative network with the partners that best fit the network's mission can be extended to non-business purposes, including incident management and disaster rescuing processes (Camarinha-Matos et al., 2005).

Whether we are talking about virtual organizations or other flexible sourcing paradigm, there is the need to develop ICT (Information and Communication Technologies) facilities that provide services and protocols for software agents to meet in a way that can be trusted, efficient and safe. In this context, several authors proposed and followed the concept of electronic institutions, i.e., frameworks that implement interactions conventions and that can offer basic services, such as interaction and negotiation protocols, ontologies, rules and norms for regulating consortia formation and subsequent joint operation, pro-active contract monitoring services, and computational trust and reputation services (e.g. Dignum and Sierra, 2001; Esteva et al., 2001; Rocha and Oliveira, 2001; Lopes Cardoso and Oliveira, 2005).

More recently, a growing focus is being given to the broader notion of *agreement technologies*, that is, technological components, processes and mechanisms that support the establishment of agreements of socially motivated agents in their interaction with others. These components can exist either in the context of electronic institutions or in any other situation where agents need to interact. In this new context, the environment where agents live has an essential role, acting not only as a facilitator (providing communication, organizational and coordination services), but also as a regulator (providing monitoring and enforcement capabilities) and a mediator, i.e., the social medium that is able to influence the agents' behavior (Oliveira, 2012).

There is currently a lot of work being done in the agreement technologies realm aiming the construction of reliable and efficient agent societies, mainly in the following generic categories:

- Privacy, security and reliability issues, including physical aspects (e.g. cryptography, safe channels, authentication) and semantic ones.
- Integration, management and planning of the business processes of all the members of the virtual community.
- Adaptive decision making in processes such as partners' selection, negotiation, argumentation, and dynamic scheduling of tasks between the members of the virtual community.

- Trust and reputation mechanisms offering trust-based services to enhance the social order between the members of the community.
- Agreement management, including the processes of creation, configuration, management and termination of agreements and contracts between the members of community.
- Normative mechanisms that exert influence and control over the members of the community.
- Standardization of procedures and technologies, and licensing.

The work presented in this thesis focus on the use of computational trust as an enabler technology for virtual societies and organizations. In the next section, we contextualize this work in the scope of a broader project being developed at the LIACC Laboratory of the University of Porto.

1.1 Contextualization

The theme of this PhD thesis came under the project Electronic Institution, being developed at the LIACC Laboratory of the University of Porto. The Electronic Institution (EI) is a computational agent-based platform that aimed to assist and coordinate the establishment of safe and reliable business agreements between agents through appropriate electronic contracting services.

When the thesis' theme was proposed, the EI was supported by two basic components: a multi-round, multi-attribute negotiation protocol (Rocha and Oliveira, 1999; Rocha et al., 2005), and an enforceable normative environment, enriched with a background normative framework facilitating contract establishment and monitoring (Lopes Cardoso, 2010; Lopes Cardoso and Oliveira, 2011). Besides these key services, the EI provided an ontology service and a notary service. It also offered a rudimentary computational trust and reputation service based on the following principles:

- The trust that an agent has on another agent under evaluation is given by a score computed using the past contracts established between both agents. In turn, the reputation of the agent under evaluation is computed using all past contracts established with this agent in the Electronic Institution.
- Trust and reputation scores must increase after the agent under evaluation fulfills an obligation and must decrease after the agent violates

an obligation. The weights ascribed to either the fulfillment or the violation of an obligation may be different.

- Different agents may have different paces in the formation of their trust on others.

In this context, it was proposed to the PhD student the development of a richer computational trust and reputation model to be integrated as a service in the EI, which would more realistically mirror the trusting process of business entities. Meanwhile, two different circumstances lead to a slight shift in the purpose of this thesis:

1. The EI platform has evolved to the ANTE (Agreement Negotiation in Normative and Trust Enabled Environments) framework, as a result of project PTDC/EIAEIA/104420/2008 sponsored by Fundação para a Ciência e a Tecnologia. As a consequence, it was considered as a desired characteristic of the resulting computational trust model that it would be applicable to other business paradigms than the virtual organization (including dyadic inter-organizational relationships), as well as to other types of social interactions (including those happening in social networks), without substantial adaptation effort.

2. After a preliminary study on trust and reputation, the work team recognized that reputation is a social phenomena *per se* as complex as trust, and addressing both concepts in the context of this thesis would prevent the deepening of the trust component. Therefore, it was decided that the PhD work should focus on trust rather than on reputation, although the resulting model for trust management should be flexible enough to account for the integration of a future reputation service.

1.2 Research Methodology

The research problem addressed in this work was originally formulated in general terms by the EI (now ANTE) project's coordination, as follows: *Current approaches to computational trust are still immature, which prevents their wide acceptance and adoption in real word electronic (social and economic) relationships. This, in turn, constitutes a barrier to the automation of business relationships in processes such as the selection of partners, assignment of tasks, and contract drafting, and to the reliance on autonomous software agents for (partial) decision making.*

In order to narrow the research problem down and rephrase it in operational terms, we conducted extensive exploratory research that allowed us to gain familiarity with the problem and achieve a thorough understanding of it. Taking into consideration the nature of the problem and the timings and resources available in the scope of the ANTE project, we proceeded with the reviewing of the literature on trust and computational trust; occasionally, we collected qualitative information through interviews to Portuguese associations and firms in the scope of FCT project PTDC/EIAEIA/104420/2008.[1]

This exploratory study allowed us to narrow the research problem as intended. We verified that most of the approaches on computational trust were not adequately grounded on the multi-disciplinary literature on trust in such aspects as the situation-awareness of trust, the consideration of different antecedents of trust and trustworthiness, and the importance of understanding trust as a social concept. Hence, we concentrated our research efforts in these sub-topics of computational trust. In the same way, due to the complexity and scope of the mentioned sub-topics, we excluded from our work other important research trends currently addressed in computation trust. Namely: the (possible) integration of reputation information in trust assessment, with particular emphasis on the credibility of this type of information; the use of new sources of information, including the argumentation about trust; and the representation of trust evidence, with particular emphasis on ontology-based representation.

The research problem at hand was then reformulated into six research questions that remain uncovered in the computational trust community and whose answer is fundamental for the credibility of computation trust models. We address these questions next.

1.2.1 Research Questions

The first question concerns the theoretical definition of trust and the way this definition is translated into a computational model. For instance, most computational trust approaches consider that trust is given by the estimated trustworthiness of the agent under evaluation that results from the aggregation of the outcomes of this agent's past actions with others. Theoretical works on trust, however, demonstrate that trust is more than trustworthiness, and hence we may wonder if a computational approach shall consider these other antecedents of trust as well in order to compute reliable estimates of the trust of an agent. A related question concerns the antecedents

[1] Lanidor, S.A., AFIA (Associação de Fabricantes para a Indústria Automóvel), Practical Way Software, and Vortal.

of trustworthiness. A given agent is generally more or less trustworthy in a given situation due to the conjugation of different factors, such as the ability of the agent in the matter in the specific situation, his integrity or even benevolence toward the agent that is evaluating him. However, most of the existent computational trust approaches does not even consider the existence of different antecedents of trustworthiness. In accordance to what was said in this paragraph, we pose the following research question:

Research Question 1 *Shall computational trust models consider other antecedents of trust than trustworthiness? In the same way, does the distinction between the ability, integrity and benevolence of an agent in a situation allow for more reliable estimations of this agent's trustworthiness?*

In the previous research question, we mentioned *en passant* the importance of the situation in trust and trustworthiness estimation. Some computational trust models propose mechanisms to infer the trustworthiness of agents in situations that are related to the current situation under evaluation, which is in fact an important question, especially in dynamic environments where the existing evidence on the agent under evaluation may be scarce. Most of those models, however, use approaches based on ontologies or similarity measures, suggesting a not so good response in these dynamic environments. Moreover, these approaches do not distinguish between the different trustworthiness antecedents, and therefore do not reason about whether each one of them is indeed situational or not. Concerning what we have just said, we pose the following research question:

Research Question 2 *Is it possible to develop situation-aware computational trust models not based on similarity distance functions that give satisfactory results even when the evidence available on the agent under evaluation is scarce?*

In trust theory, there are several references to the dynamics of trust building. For example, the weight of a deceptive event may be stronger than the weight of a positive event, and both weights may be different depending on the trust relationship that exists between the agent that trusts and the one that is trusted. The stage of the relationship between both agents, as well as its evolution with time and situation, is thus very important in trust assessment. However, the existent computational trust approaches do not consider or use the benevolence of agents toward specific partners in trust judgments. This lead us to an additional research question, as follows:

Research Question 3 *Does the extraction of the benevolence state of a given agent toward his evaluator from the available evidence improve the reliability of trust judgments?*

A side question concerns the evaluation of the existent computational trust approaches. Taking into consideration that trust has its own dynamics and that the relation between the one that trusts and the one that is trusted is paramount to understand past trusting decisions, we firmly believe that the current experimental approaches used to evaluate computational trust models, which are essentially based on simple distributions modeling the behavior of agents, need to be reformulated in order to capture the evolving behavior of agents and improve the acceptability of the validation of these models. Based on that, we pose the following question:

Research Question 4 *Is it possible to build a model of behavior of interacting partners that is simple enough to allow for the fair comparison between different computational trust approaches and still allows for realistic and relational evolving behaviors?*

A different question concerns our belief that some of the existent computational trust models work better when there are several individuals items of evidence about the past performance of the agents under evaluation. However, there are several environments where trust decisions have to be taken even when the feedback about the agents under evaluation is scarce. For instance, Wathne and Heide (2000) cited an example of the diamond trade in New York, dominated by a close-knit community of traders, where information about opportunistic behavior spreads rapidly; in this case, opportunists are rapidly excluded and trust decisions are made even before a long set of evidence is available about the agents' behavior. In this context, we formulate the following research question:

Research Question 5 *How to build robust and reliable trustworthiness estimators that work in a satisfactory manner even when the number of individual items of evidence on the agent under evaluation is small?*

Finally, we consider the fact that trust is seen as a fundamental factor to social order. However, not all interactions between individuals and/or collectives happen within the context of trust relationships; in certain cases, other means are necessary to secure these interactions, such as incentives and social control in the form of norms and sanctions, but these means are frequently costly. Ideally, the interplay between trust, norms and sanctions

that exists in society should be transposed to artificial societies. This lead us to our final research question, as follows:

Research Question 6 *How to transpose to artificial societies the complementarity that exists in our society between trust, norms and sanctions?*

1.2.2 Research Process and Methods

After redefining the research problem, we proceeded with the common steps of the research process to address the formulated problem. The main methods and techniques we adopted were chosen taking into consideration the applied (problem-oriented) nature of this research. By this, we mean that we are not doing pure research, in the sense that our aim is not just the formalization of a theoretical model. On the contrary, we are pursuing what is sometimes named by 'instrumentalist' kind of research, contributing to making human intervention in the real world environments more effective. Following the instrumentalist terminology, which considers two major alternatives of research (i.e., applied vs. problem-oriented), we consider our work to be problem-oriented. This means that we did not start from known techniques and tried to choose the best ones to apply but, instead, we started from the problem formulation and tried to derive an appropriate model and process to cope with it.

We conducted another phase of intense background research through the study of the relevant literature on trust and social trust, from a multidisciplinary perspective, aiming to achieve new insights on the key concepts associated to each one of the research questions. The main outcomes of this study are presented in Chapter 2 of this thesis. At the end of this study, we analyzed how these concepts could be transposed to a computational trust model, and reviewed the existent computational trust models looking for particular implementations of the concepts; when such models existed, we analyzed them in more detail in order to verify if they adequately fitted the theoretical concept (Chapter 3).

After this first phase of exploratory research, we developed the working hypotheses that guided our research throughout the remaining phases of our work (cf. Section 4.1). We planned to test those hypotheses empirically through simulation, in an artificial environment within which relevant data could be generated. For this, we developed two distinct computational models. The first of these models consisted in a generic framework for situation-, benevolence-based social trust proposed by us that we instantiated into a

computational prototype. The main components of the resulting computational trust model are presented in Chapter 4. The second model emulated the behavior of individuals (or collectives) with different characteristics of ability, integrity and benevolence that establish different benevolent relations based on these characteristics and on mutualistic interests (cf. Section 5.4). This model was based on postulates grounded on the theory of trust, benevolence, and integrity and was also inspired by real data resulting from inquiries made to 127 Portuguese firms selected from the SABI (Iberian Balances Analysis System) firms database, in the scope of the PTDC/EIA-EIA/104420/2008 project (cf. Alves et al., 2012).[2]

After the creation of the simulation testbed, we performed the experiments that enabled us to collect the data through the observation of quantitative measurements. Then, we proceed to the quantitative analysis of data, where we compared the results obtained with our computational trust model with those obtained with other computational trust models that did not presented the situation-aware and social-based features that we were testing, for the observed measurements. We conducted two-sample t-tests to decide about the significance of the obtained results. Finally, we draw and reported our conclusions about the validity of the constructed hypothesis. The steps associated with data collection and analysis are further described in Chapter 5. The final conclusions about the work developed in this PhD thesis are summarized in Chapter 7.

1.3 Contributions

The work of this thesis has three main contributions, that we describe next. The first contribution is the SOLUM (Situational-aware and sOcial computationaL trUst Model) framework, a generic framework for social trust, composed of five distinct evaluation functions, which may be instantiated in different models of social-based computational trust. The key elements of this framework were designed taking as a starting point the insights from our thorough multi-disciplinary study on trust as a social construct. They reflect the key topics of our research, such as the estimation of the ability, integrity, and benevolence of agents based on the evidence available on these agents, the current situation, and possibly the agents' reputation; and also

[2]It is worth to note that we considered the use of real data to test the hypotheses; however, this type of data is very hard to obtain in practice. In the same way, the validation of our computational trust model in real scenarios (e.g., within a firm conducting sourcing activities) that would allow for the hypotheses testing was out of the scope of the FCT project supporting this research.

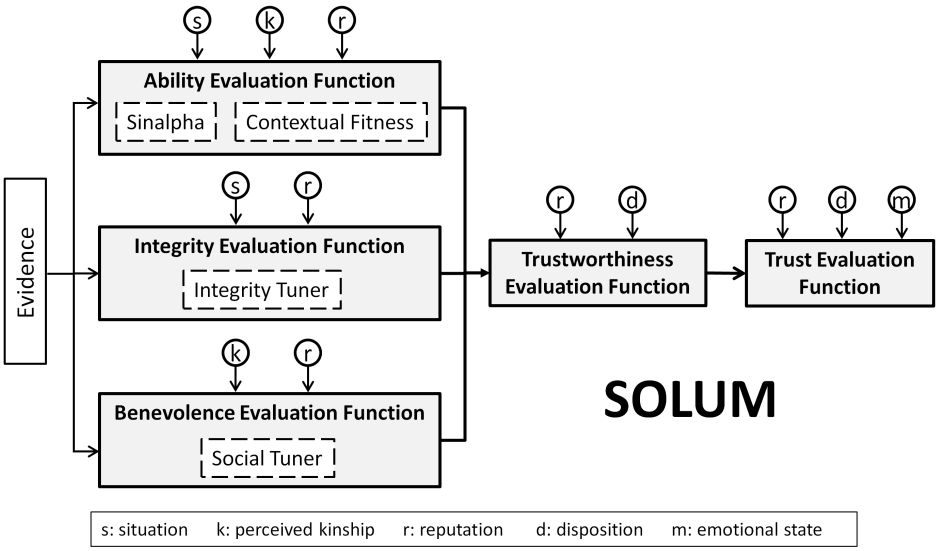

Figure 1.1: The SOLUM model.

the consideration of the disposition and emotional states of agents as other antecedents of trustworthiness and/or trust. The presentation of this study in Chapter 2 is, itself, a side contribution of this thesis. A high-level diagram of SOLUM is given in Figure 1.1, where the main elements of the framework are represented in solid line.

The second contribution of our work is the creation of four different computational trust components used to test the hypotheses derived from the research questions enunciated in Section 1.2, and which constitute our proposal to instantiate the above mentioned framework. These components, which are represented in Figure 1.1 in dashed lines, are:

1. *Sinalpha*, an estimator of the overall trustworthiness of the agent under evaluation. This component takes into consideration the fact that the perceived trustworthiness of an agent depends on the specific stage of the relationship between this agent and the one that is evaluating him.

2. *Contextual Fitness*, a component that analyzes the past evidence on the agent under evaluation and detects tendencies on his behavior in particular situations. Hence, it is a component that allows any generic trustworthiness estimator to become situation-aware. The core of this component is the information gain metric (Quinlan, 1986), which we use here as an online process, making it well suited to open and dynamic environments where the available evidence on any agent may be scarce.

3. *Social Tuner*, a novel trust-based component that analyzes the past interactions between the one that trusts and the agent under evaluation and estimates the benevolence of the latter toward the former, from the outcome of these interactions.

4. *Integrity Tuner*, a novel trust-based component that analyzes all past evidence on the agent under evaluation and estimates his overall integrity, as related to the consistency of his actions and his ethics.

We opted to develop such components as separate and independent modules, allowing them to be added to any existent computational trust model. Another important component of our model is our instantiation of the trustworthiness evaluation function (cf. Figure 1.1), which is able to weight the contribution of each one of the described components in the final trustworthiness score based on the stage of the relationship between the agent under evaluation and his evaluator. We think that this is an important advance over the state-of-the-art on computational trust, in the sense that very few computation trust models address integrity and benevolence as antecedents of trustworthiness, distinguishing them from the general competence of the agent under evaluation, and, to the best of our knowledge, only the Socio-Cognitive Model of Trust (Castelfranchi and Falcone, 2010) implements such a distinction. However, we go one step further when we propose to weight each antecedent based on the *stage of the relationship* between the interaction partners. This is an important aspect of the so called social trust, and the field of computational trust is now eagerly looking for computational versions of social trust.

Finally, the third main contribution of our work lies on the analysis that we conducted about the interplay between trust, norms and sanctions. In order to perform the empirical part of this analysis, we adapted the normative environment of the ANTE framework (Lopes Cardoso, 2010; Lopes Cardoso et al., 2012) to allow the automatic drafting of contractual sanctions based on the agents' estimated trustworthiness, as well as to allow the generation of trust-based evidence from the facts that resulted from monitoring the established contracts.[3] The conjugate use of both forms of social control has received some attention from theoretical literature, but practical implementations of this interplay are rare, despite its well-accepted relevance to the development of reliable and efficient agreement technologies. We believe that our work provide valuable insights into how further work on the matter may be developed.

[3]This adaption was performed in close collaboration with all the elements of the ANTE project.

1.4 Thesis Structure

The rest of the thesis is structured as follows:

- Chapter 2 provides an overview of social trust, paying particular attention to the difference that exists between this construct and trustworthiness, and to the different perspectives of trust, including its factors, nature, and dynamics. The importance of the relationship that exists between the one that trusts and the one that is trusted becomes evident throughout the chapter.

- Chapter 3 gives a review of existing computational trust approaches, more particularly how they cover the key concepts of social trust discussed in the previous chapter. We show that most of these concepts are partially addressed by some of these models, but few of them address the distinction between trust and trustworthiness or distinguish between the trustworthiness factors or dimensions. In reality, we show that the great majority of the existing computational models of trust fail to capture the distinct perspectives of social trust.

- Chapter 4 introduces the SOLUM model, which integrates the generic SOLUM framework and our current instantiation of this framework, composed by the *Sinalpha*, *Contextual Fitness*, *Social Tuner* and *Integrity Tuner* components.

- Chapter 5 presents the evaluation of the SOLUM model. This chapter presents unit tests of *Contextual Fitness* and *Social Tuner*, as well as a more elaborated evaluation of the SOLUM model, using dynamic and evolving models of agents' behavior.

- Chapter 6 describes how our computational trust model is used as a service in the ANTE platform, and how it interfaces with the negotiation and normative environment services. We also present the experimental analysis we have made to the interplay between trust, norms and sanctions.

- Finally, Chapter 7 summarizes the main results we have obtained in this work. In the same way, the main limitations and virtues of this work are pinpointed, and our plans for future work are described.

Chapter 2

A Global View of Social Trust

Trust is a very complex concept to work with. It is omnipresent in our lives, it is said to fuel social relationships and, yet, humans generally do not consciously model it in day life. Trust is ambiguous and illusive both is meaning and usage. As put by Hardin (2004), most research on trust is based on vernacular use of the term, which is, itself, a term of the vernacular. And it is in the domain of the vernacular that we propose to make a preamble to this chapter by analyzing an example of the ambiguity and complexity of the notion of trust, provided by the following excerpt of the lyrics of an old song by the pop band Genesis:[1]

I need someone to believe in, someone to trust.
I'd rather trust a countryman than a townman,
You can judge by his eyes, take a look if you can,
He'll smile through his guard,
Survival trains hard.
I'd rather trust a man who works with his hands
He looks at you once, you know he understands,
Don't need any shield,
When you're out in the field.

These lyrics are apparently simple to understand and it is not evident at first sight why one should even care to analyze them. The fictional character uttering such words, which may or may be not the alter ego of the author of the lyrics, trusts simple people that work in the field. We have then strong

[1]The Chamber of 32 Doors (from the album The Lamb Lies Down on Broadway, 1974).

evidence on the narrator trusting country working people. If someone asks our opinion, we would say that the narrator trusts country people because he told us so. And finally, if we would feed a computational trust system in order to estimate the trust that the narrator put on some field worker, certainly that this utterance would be weighted as strong evidence.

However, if we care to have a deeper thought about the lyrics, some improbable questions would raise. On the one hand, we can question the reasons underlying the strong opinion of the narrator concerning trusting country people. Are they motivated by some dispositional, psychological characteristics of him? Are they the result of a cognitive process resulting from the interaction of the narrator with the environment through his lifetime? Is the narrator playing rational over the possible outcomes, hidden in the uttered words? On the other hand, we can question the scope (and the truth) of the illocutionary act itself. In fact, it is not probable, using common sense, that someone blindly trusts a whole category of people (or institutions or things) across all possible situations. Probably the narrator trusts more *some* country man than other. Probably there is some town man for which the narrator put more trust than a specific country man. Probably our fictional character trusts generally the country man but would trust more a town man to make his (hypothetical) daughter happy in marriage. And probably (as we do not know!), the narrator likes to think he trusts the country man because of political ideals. Finally, when making a trust-based decision about who to risk an interaction in a practical situation, it would be of no surprise if the narrator choose the town man to interact with.

With this small introductory example, we observed several different aspects of trust showing up confusingly, such as the situationality of trust and the notion of trust as knowledge or act. As we are going to observe throughout this chapter, this fictionary example reflects the current state of research on trust, where different theories appear sometimes to be in opposition, others simply cannot be included in the same theoretically framework. However, it is not possible to build any coherent computational trust model without knowing and understanding the basic principles of the social phenomenon. In our point of view, so many existent computational trust proposals fail to model trust because they are not grounded on an extensive and exhaustive study of trust as a social concept. Therefore, this chapter presents a detailed description of the theory of social trust. Whenever possible, we capture the theoretical aspects of trust as simple propositions that help to systematize the discussed concepts.

It is important to clarify that not all of the concepts that we address in this chapter are easily translated into a computational model of trust.

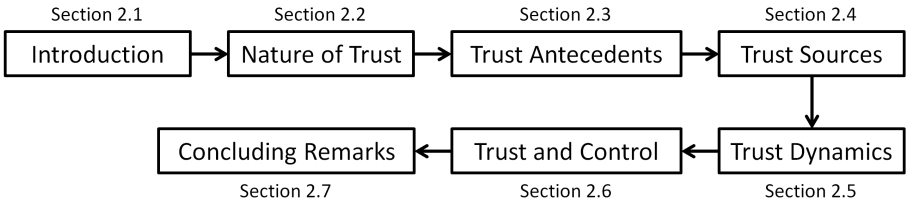

Figure 2.1: A guide for this chapter.

In fact, our model of computational trust that we present in later chapters does not implement all the propositions that we derive here. However, by having this broader picture of trust, we have a relative safe guarantee that what we transpose to the computational model is grounded on trust theory.

Because we address a wide range of topics, Figure 2.1 provides a roadmap to the contents of this chapter. We start by presenting generic and introductory aspects of trust. Then, in Section 2.2, we focus on the nature of trust, including the different accounts referred to in literature (cognitive, emotional, and behavioral) and the situation-awareness of trust. Section 2.3 addresses a question of paramount importance to computational trust: which are the antecedents of trust, i. e., based on what an individual forms his trust on a specific object of trust? A related question concerning the different sources of information that can provide information for trust reasoning is addressed in Section 2.4. In Section 2.5, we overview the dynamics of trust creation and maintenance, addressing relevant topics such as how trust forms, how is the path that the object of trust have to run in order to be considered highly trustworthy, and what is the effect of betrayals on trust. Section 2.6 is devoted to the relation between trust other forms of social control, such as contractual norms. Not only this is a relevant topic on the study of trust by itself, but also it is fundamental for the work we are doing at LIACC in the aim of our funded research project. Finally, Section 2.7 presents the concluding remarks for the section and refers advanced topics on trust that were not overviewed in this chapter.

2.1 Introduction

Trust is a social construct that is present in the day-to-day routine of humans. In fact, every time a person (hereafter named *truster*[2]) needs to interact with, delegate to or rely on an action of another individual, group or thing (hereafter named *trustee*), a decision about trust is made.

[2]Some authors use instead the word *trustor*, and some others even *trustier*.

Trust is of particular interest in social and economic exchange relationships.[3] Both, in a way, imply uncertainty, normally associated to information asymmetry regarding the attributes and actions of the partner (Wathne and Heide, 2000), which, in turn, increase the vulnerability of the actors that engage in particular interactions (Heimer, 2001). Although the definition of trust varies throughout the trust literature, we start this section with the definition presented in Mayer et al. (1995), where trust is both a positive expectation that the trustee will act as expected, and a willingness or intention to accept vulnerability.

In the same vein, uncertainty and vulnerability increase the risk of opportunism between the interacting partners (Wathne and Heide, 2000). In order to reduce uncertainty and vulnerability, different mechanisms may be used, such as control, monitoring, and incentives (Wathne and Heide, 2000). However, these mechanisms are generally costly and some authors propose the use of trust as an alternative governance structure (Sako, 2002).

Hence, due to the vital role that trust plays in the society – and to the disturbing effect of its counterpart, *distrust* –, it is of no surprise that it has been receiving increased attention from researchers in several areas, including sociology, economics, management, political science, psychology, ethics and philosophy (e.g. Macy and Skvoretz, 1998; Dasgupta, 2000; Hardin, 2001; Kiyonari et al., 2006; Wathne and Heide, 2000; Heimer, 2001; Castelfranchi and Falcone, 2010). More recently, trust management started receiving growing attention from the computer science community, particularly from multi-agent systems scholars. The underlying idea is to confer to intelligent agents the ability to estimate the trustworthiness of their interacting partners, in order to improve their social interactions (Sabater-Mir and Paolucci, 2007). In this case, we say that agents use *computational trust models* based on trust theory to assist their trust-based decisions.

There are several different approaches to the study of trust and its dynamics and distinct discussions of this social concept are still ongoing. Some approaches, mostly in psychology, consider trust at the individual level, as personality traits or dispositions of the truster that develop at the infancy and remains stable through adulthood (e.g. Rotter, 1967; Cvetkovich et al., 2002; Kiyonari et al., 2006). Some of these studies use questionnaire surveys and other psychometric scaling techniques that ask participants to evaluate their trust on some entity after some suggested event. Others study the

[3]Following Blau, cited by Colquitt et al. (2007), economic exchanges are contractual in nature and concerns the exchange of quantities agreed upon beforehand, while social exchange are more diffuse in nature, vaguely defining future obligations that occur over a more open-ended time frame.

behavioral expressions of trust in laboratory settings, using different kinds of trust games, such as the Trust Game, the Dictator Game, the Investment Game and the Faith Game.[4] General criticism about these approaches concerns the fact that trust cannot be reduced to a trait-like notion, and that individuals would have no need to trust apart from social relationships, so that trust must be understood instead as a multidimensional social reality (Lewis and Weigert, 1985; Mayer et al., 1995; Schoorman et al., 2007). Other criticism concerns the methodology used in these studies, where the data resulting from the studies is generally obtained through participation in group activities and do not capture other psychology data than "correlations between measures over time of generalized trust" (Hardin, 2001).

Other approaches, mostly in the area of social psychology and sociology, study the role of cognition, affection, and values in trust formation and maintenance, in the context of social relationships (e.g., Fitness, 2001; Finkel et al., 2002; Hardin, 2004). This means that trust must be understood as a property of ongoing dyads, groups, and collectivities, and not of isolated individuals (Lewis and Weigert, 1985; Mayer et al., 1995). The sociological view of trust expands to demonstrate the fundamental role of trust across the different social institutions (e.g., Luhmann, 1979; Lewis and Weigert, 1985). Still other approaches assume an economic focus where trust is affected by situational factors, such as incentives, norms, institutions, and other governance mechanisms (e.g. Dasgupta, 2000; Williamson, 1979; Ireland and Webb, 2007; Wathne and Heide, 2000; Sako, 2002). In managerial and organizational research, there is important studies on opportunism and betrayal within and between organizations (e.g., Elangovan and Shapiro, 1998; Wathne and Heide, 2000; Williamson, 1979).

Such a diversity of notions and concepts – where scholars from different disciplines seem to "talk past one another" (Schoorman et al., 2007) – derives from the fact that "trust is somewhat illusive, difficult to define (...) and difficult to measure" (Elofson, 1998). It reveals, in the words of Castelfranchi and Falcone (2010), a "degree of confusion and ambiguity that plagues current definitions of trust". This by no means eases the work of computer scientists when they attempt to formalize models of computational trust for assisting the decision making of artificial entities.

A frequent misconception on trust literature concerns the distinction between trust and trustworthiness. This way, we start our study on trust by analyzing the main differences between both concepts and the way they relate. As we are going to see, they can be decomposed in distinct dimensions,

[4]See Kiyonari et al. (2006) for a description of these games.

for which contribute different factors. Still in this section, we analyze the important interrelation between trust and context. At the end of the section, we propose a tentative formalization of the trust concept, taking into consideration its late use in computational models.

2.2 The Nature of Trust

Some time ago, the idea of generalized trust associated with the propensity to trust of humans was popular among some scholars: some individuals were considered to be high trusters and others low trusters. This idea is being abandoned, and there is now a generalized consensus within the trust community that trust is – at least – a ternary relation: the truster trusts the trustee with respect to some matter (Hardin, 2001). In fact, there may be different levels of trust even in the same relationship, not only because the trustee may show different qualities in different domains (Mayer et al., 1995; Kelton et al., 2008), but also because the trust requirements of the truster may change within the relationship. For example, an agent may trust a colleague to do a good job collaborating on a research project but not teaching his class in his absence (Schoorman et al., 2007). A much more evident example is given by Marsh (1994), when he says that an agent may trust his brother to drive him to the airport, but not to fly the plane.

Another characteristic of trust is that it is not necessarily mutual: the truster may trust the trustee, but the latter may not trust the former (Schoorman et al., 2007).

2.2.1 The Situationality of Trust

A consensual notion in the studies on trust is that it is *situational*, in the sense that persons have different incentives, competences and abilities to be trustworthy on different occasions (Marsh, 1994; Dasgupta, 2000; Dimitrakos, 2002). For instance, a person can trust his neighbor to have the key of his house in "normal circumstances" but stop trusting if the neighbor is detected an alcoholic problem; or I can trust my friend to drive in hot summer but not in snowing conditions; my trust on my business partner may diminished if he is having problems with his logistics.

Proposition 1 Trust as a quaternary relation*: Trust is a property of the truster in relation to the trustee with respect to some matter and in a given context.*

2.2.2 The Cognitive, Emotional and Behavioral Accounts of Trust

Trust is a cognitive process. It is an assessment of the trustworthiness of the object of trust, being it a person, group, or institute, and this assessment is based on evidence; it is up to individuals to choose whom to trust, and in which circumstances (Lewis and Weigert, 1985; Mayer et al., 1995; Hardin, 2004; Castelfranchi and Falcone, 2010). Castelfranchi and Falcone, followed by Herzig et al. (2010), consider that to trust implies to have a goal, which can be accomplished by an action of the trustee, and that trust forms in a complex cognitive construction based on beliefs and meta-beliefs about the trustee (Castelfranchi and Falcone, 1998; Castelfranchi et al., 2003; Castelfranchi and Falcone, 2010).

Complementary to its cognitive base, trust is also constructed on an emotional base. On the one hand, participants in ongoing close relationships create an emotional bond between them (Lewis and Weigert, 1985; Mayer et al., 1995; Castelfranchi and Falcone, 2010); on the other hand, emotional states, even when unrelated to the trustee or situation, may affect trust (Dunn and Schweitzer, 2005; Schoorman et al., 2007). These emotions influence the mental attitude and perceptions of the relevant virtues that are needed for relying on the trustee (Schoorman et al., 2007; Castelfranchi and Falcone, 2010).

Proposition 2 Cognitive and emotional content of trust*: Trust is a cognitive process that involves the estimation of the trustworthiness of a trustee based on the evidence available on the trustee. Complementary to its cognitive content, trust has an emotional content, which is particularly relevant in close and ongoing relationships and strong emotional states.*

Schoorman et al. (2007) refer that emotions may create a temporary 'irrationality' about the cognitive assessment of ability, integrity and benevolence, but do not discriminate in which ways each one of these antecedents of trust are affected. Moreover, these authors suggest that after a violation of trust the perceptions may gradually return to a rational perspective, although they question if the emotional account of the evaluation completely dissipates or in some way remains with time. We think that the authors refer to the dissipation of affect after a betrayal; as we are going to see later in this chapter, other emotions (e.g., sadness, rage) raise after a betrayal, and, in our opinion, these other emotions will also create a temporary 'irrationality' about the cognitive assessment of the above mentioned antecedents of trustworthiness.

Proposition 3 Effect of Positive Emotions on Trust*: Affect and bond creation have a positive influence on the perception of the antecedents of trust. This influence dissipates after a (mild) violation of trust.*

Apart from the cognitive and emotional content of trust, some scholars consider that trust includes the behavioral enactment of trust (Luhmann, 1979; Lewis and Weigert, 1985; Castelfranchi and Falcone, 2010), which means that trust is also a decision and an act of relying on, counting on, depending on the trustee, and that the cognitive, emotional, and behavioral dimensions are united over a common structure. Other authors, however, refute this behavioral account of trust, referring that trust is just a piece of knowledge, which can eventually be translated into a willingness to take risk (e.g. Mayer et al., 1995; Hardin, 2001), while trusting behavior is the actual assuming of risk.

Proposition 4 Behavioral content of trust*: Some authors consider that trust has a behavioral content; some others consider that the notion of trust does not include the decision and act of relying on the trustee.*

2.2.3 The Degree of Trust

In the same way, trust is a matter of degree (e.g. Hardin, 2001; Bhattacharya et al., 1998). As mentioned by Castelfranchi and Falcone, "only a trust decision eventually is a yes/no choice, and clearly needs some threshold" (Castelfranchi and Falcone, 2010). The authors introduce the degree of believing as the basis for the degree of trust.

Trust has a strength related to the confidence that the truster has on his trust (e.g. Bhattacharya et al., 1998; Huynh et al., 2006; Patel, 2006).

Proposition 5 Degree of trust*: Trust is a matter of degree.*

2.3 Trust and Its Antecedents

In order to model trust, the way it forms and its dynamics, we need to identify and characterize its antecedents. This is an underrated question in computational trust, where most of the existing approaches (with important exceptions, that we identify later) model trust as the direct reflex of the trustee's trustworthiness. [5]

[5]The outcomes of trust, such as risk taking and job performance, are out of the scope of this thesis (Mayer et al., 1995; Colquitt et al., 2007).

Although trustworthiness is an important antecedent of trust, it does not explain all trust dispositions and decisions. Other factors must be considered, such as the propensity to trust of the truster (Mayer et al., 1995; Kiyonari et al., 2006; Colquitt et al., 2007; Castelfranchi and Falcone, 2010), his emotional state (Schoorman et al., 2007), the physical and cultural characteristics of the trustee (Kelton et al., 2008), and (for some scholars) reputation (e.g., Jøsang and Ismail, 2002; Kelton et al., 2008; Castelfranchi and Falcone, 2010). We analyze each one of these antecedents with more detail next.

Proposition 6 Antecedents to Trust*: The trustee's trustworthiness, his physical and cultural characteristics, and the truster's propensity to trust and his emotional state, all are antecedents to trust. Some authors also refer reputation as another antecedent to trust.*

2.3.1 Trustworthiness

Trust and trustworthiness are two distinct concepts: trust is a property of the truster in relation to the trustee, while trustworthiness is a characteristic of the latter (e.g. Hardin, 2002; Kiyonari et al., 2006; Colquitt et al., 2007; Castelfranchi and Falcone, 2010), i.e., a multifaceted construct that captures the trustee's competence and character (Colquitt et al., 2007). A trustworthy entity is the one worthy of confidence, it normally would present high values of competence, integrity and benevolence in the situation in assessment, and its behavior would be predictable in this situation.[6] To paraphrase Hardin (2004), "if, on your own knowledge, I seem to be trustworthy to some degree with respect to some matter, then you trust me with respect to that matter".

Proposition 7 Trustworthiness as Multi-Dimensional*: Trustworthiness is a multi-dimensional construct that captures the trustee's competence and character.*

The trustworthiness of the trustee in a given situation is objective; however, the trusting agent may make a wrong evaluation of the trustee's reliability – possibly by conducting an insufficient or deficient gathering of data on the trustee –, which may lead to a misplacement of the truster's trust. This means that trusting agents deal with the *perceived* or evaluated trustworthiness, which is subjective (Castelfranchi and Falcone, 2010).

[6]Schoorman et al. (2007) apply this definition to either interpersonal, intergroup, and interorganizational levels of analysis.

Next, we take a closer look at the three dimensions (or factors) of trustworthiness proposed by Mayer, Davis, and Schoorman (1995), and followed by other scholars (e.g. Elangovan and Shapiro, 1998; Colquitt et al., 2007): ability, benevolence, and integrity.

Ability

Ability, also referred to as *competence*, relates to the potential ability of the evaluated entity to do a given task, and is one of trustworthiness dimension most mentioned by trust scholars (e.g. Mayer et al., 1995; Hardin, 2002; Levin et al., 2004; Xie and Peng, 2009; Castelfranchi and Falcone, 2010; Adali et al., 2011). It translates into a set of qualities that makes the trustee able for the task, such as skills, know how, expertise knowledge, general wisdom, self-esteem, interpersonal skills, self-confidence, and leadership. A trustee that shows some or all of these qualities is contributing for the truster's perception that he has the ability to perform the task. The perception of these attributes by the truster is mainly a cognitive process and less of an emotion-based process (Colquitt et al., 2007). Also, ability is domain specific; for example, a given trustee can master some technical issue and still show little aptitude in interpersonal communication (Mayer et al., 1995).

Certain attributes are sensitive to context. For instance, Hardin (2002) refers that experience and age are desired characteristics in babysitting activities. In the marketing area, Xie and Peng (2009) show that informational recovery efforts (e.g. presenting sufficient or persuasive information about the events) in trust repair activities of a firm (e.g. following negative publicity) improve perceptions of organizational competence. In general, we can state that the competence dimension is a major issue in contexts where specialized abilities are at issue (Hardin, 2002).

Proposition 8 Ability*: Individuals have different abilities in performing different tasks, which are related to their inherent qualities.*

Proposition 9 Perception of Ability*: Individuals that show the qualities required to perform the task at hand are perceived by others as having ability in that matter. Moreover, the physical and cultural characteristics manifested by these individuals may increase or decrease the perception of their ability in the matter by others.*

Benevolence

Benevolence is considered by several scholars as a key element of close relationships and trust relations (e.g. Mayer et al., 1995; Elangovan and Shapiro,

1998; Lee et al., 2004; Levin et al., 2004; Platek et al., 2009; Koscik and Tranel, 2011). When assessing the trustworthiness of an interacting partner, it is important to understand if this partner is behaving, or is estimated to behave, in a benevolent way in the particular relationship. Therefore, benevolence must be correctly understood and modeled.

Proposition 10 Importance of Benevolence*: When assessing the trustworthiness of an individual, it is important to estimate his benevolence toward the truster.*

The Merriam-Webster.com dictionary defines benevolence as both a *disposition to do good* and *an act of kindness*. This is consistent with some academic literature that relate benevolence to a feeling of goodwill toward the interacting partner (Elangovan and Shapiro, 1998), excluding any intention of harming him given the opportunity to do so (Levin et al., 2004). In some way, it can also mean the positive intentions referred to by Adali et al. (2011) in their trust model.

Some authors consider that benevolence implies a specific attachment of the truster toward the trusted one, involving a truly sense of loyalty and caring, and exclude from its definition any motivation based on egocentric profit motives (e.g. Mayer et al., 1995; Elangovan and Shapiro, 1998). For instance, people tend to act benevolently toward victims of an unpredictable accident, with no apparent self-benefit. In this case, it seems plausible that benevolence does not necessarily involve symmetric relations between both agents (Castelfranchi and Falcone, 2010).

Individuals have a disposition toward benevolence. In fact, some scholars (e.g., McCullough and Hoyt (2002); Roccas et al. (2002)) link benevolence to the Agreeableness and Neuroticism personality traits of the Big Five model[7], with Agreeableness being negatively correlated with benevolence. On the other hand, studies in human behavioral genetics refer that Neuroticism and Agreeableness are influenced by both heredity and environment, with a slight prevalence of the environment factor (e.g., Bouchard and McGue (2003)), and that Agreeableness tend to increase with time, while Neuroticism tend to decrease with time in women and to remain stable among man (e.g., Srivastava et al. (2003)).

Recent studies in the area of behavioral neurology and cognitive neuroscience try to uncover the role of human amygdala in expressing benevolence

[7]The Big Five or Five Factors model is a framework of personality traits, whose main traits or factors are Openness, Extraversion, Conscientiousness, Neuroticism, and Agreeableness (Roccas et al. (2002)). An individual with a strong value of Agreeableness tend to be compassionate and cooperative with others. An individual with high values of Neuroticism tend to have negative emotions.

and normal interpersonal trust. Koscik and Tranel (2011) use a multi-round version of the Trust Game to test situations involving interpersonal trust (whether or not the opponent will return a profit on an investment) and reciprocation (whether or not the opponent will betray trust). They conclude that individuals with unilateral damage to the amygdala tended to behave in a benevolent way and to increase trust in response to betrayals. In opposition, neurologically normal adults tended to act in a 'Tit-for-Tat' manner, by decreasing interpersonal trust in response to betrayals and rewarding expressions of trust.

It seems plausible that some individuals may be more benevolent than others in identical situations.

Proposition 11 Disposition to Benevolence*: Each individual has a specific disposition to benevolence, related with his traits of personality.*

Other studies suggest the involvement of the amygdala in extracting trustworthiness information from faces (Platek et al., 2009; Bzdok et al., 2011, e.g.,). In one of these studies, Platek, Krill, and Wilson (2009) use implicit trustworthiness ratings for self-resembling faces and their findings suggest that humans have evolved to use neural mechanisms that drive prosocial behavior toward kin. This means that individuals tend to be more benevolent with self-resembling others. This idea is reinforced by studies in sociobiology that suggest that there is a genetic predisposition for altruism toward close genetic kin that can overcome selection pressures favoring self-interested behaviors (Allison, 1992).[8]

Allison (1992) proposes that, in analogy to what happens with the generic predisposition for altruism, beneficent behavior toward non kin can be explained by beneficent norms that evolve based on cultural relatedness. Levin et al. (2004) refer that in the knowledge sharing domain, common language and common vision enhance benevolence between co-workers. Foddy et al. (2009) describe in-group awareness as an antecedent of benevolence. Lee et al. (2008) refer that in importer/exporter relationships cultural distance negatively impacts social satisfaction. Hence, all this indicates that the perception of kinship (e.g., through self-face resemblance) and cultural relatedness might play an important role in benevolence formation.

Proposition 12 Benevolence With Alike*: The perception of kinship and/or cultural relatedness increases the benevolence of individuals.*

[8]Although altruism reduces the reproductive fitness of the donor, performing an altruistic act toward a close relative would increase the fitness of the relative, who has a high probability of carrying the same altruistic gene (Allison, 1992).

Benevolence also develops in long-term and close relationships. In this relations, trust is reciprocated, and positive affect circulates among those who express trust behaviorally, which may result in intense emotional investments being made (Lewis and Weigert (1985)). Hence, social satisfaction lead to affective commitment, which in turn has a positive impact on the benevolence of an individual toward his partner in the relationship. The partners to the exchange develop a strong sense of loyalty and belongingness to the relationship *per se*, where relational norms and values are internalized (Lee et al., 2008).

Proposition 13 Relational Benevolence*: In long-term and close relationships, affective commitment arise and has a positive impact on the benevolence of partners.*

Some authors consider that there is a different type of benevolence that is motivated by the expectation of joint gain (Allison, 1992; Lee et al., 2008), where the voluntary helping behaviors beyond the call of duty still exists.[9] This view is consistent with Hardin (2000)'s notion of encapsulated interest, where an individual has the incentive to cooperate with his peer if he feels that taking the other's interest into account makes the peer to also take his interests into account. This is particularly true in ongoing relationships involving regularly exchanges between both partners. Hence, the partners perform acts of benevolence, which eventually may lead to true relational benevolence. We name this form of benevolence as mutualistic benevolence, using Lee et al. (2008)'s terminology, and we draw on their work to suggest the relation between mutulistic benevolence and benevolence, as illustrated in Figure 2.2.

More on Mutualistic Benevolence. Partners generally benefit from establishing ongoing relationships of trust and trustworthiness where they regularly exchange with each other over some range of matters (Hardin, 2000). In these relationships, they share a general expectancy of iterated interactions and have the incentive to cooperate because they seek the benefits associated to long-term relationships, such as the engagement in open exchange of ideas and joint learning (Elangovan and Shapiro (1998); Ireland and Webb (2007)), the expectancy that short-term inequities are resolved easily and amicably (Elangovan and Shapiro (1998)), and the sharing of risks

[9] Lee et al. (2008) refer to the benevolence based on the expectation of mutual gain in the long run as *mutualistic benevolence*, in opposition to the *altruistic benevolence* that is based on an altruistic motive. However, it is important to note that altruism requires self-sacrifice, which is rejected in most definitions of benevolence.

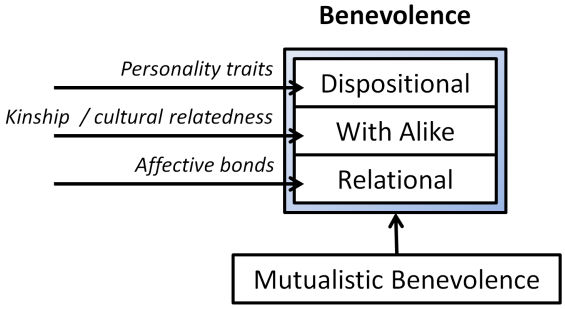

Figure 2.2: Formation of benevolence.

and costs when exploring new opportunities (Ireland and Webb (2007)). Some scholars in organizational research consider that when partners are willing to act in ways that exceed what was agreed before, the goodwill form of trust arises (e.g., Sako (1998); Ireland and Webb (2007)).

Hence, ongoing interactions increase the performance of partners (Lee et al., 2008) and their satisfaction with the relationship; as a consequence, the partners tend to act benevolently with each other (Elangovan and Shapiro (1998); Hardin (2000); Ireland and Webb (2007)). Conversely, the situational satisfaction of partners lowers down if they perceive the end of the relationship (Elangovan and Shapiro (1998)). Research studies on close relationships and deceit refer that the perceived equity of exchange between the parties is another factor that significantly influences the assessment of a relationship (Elangovan and Shapiro (1998); Lee et al. (2008)), as the perception of any inequity downgrades the value of the relationship (Elangovan and Shapiro (1998)). If the partner envisions the end of future exchanges or is in the presence of an unanticipated contingency, he may have no more incentives to cooperation, and may stop acting benevolently (Ireland and Webb, 2007).

Additionally, the value that an individual attaches to a given trust relationship may diminish if he perceives that the likelihood of being trusted by somebody else is high (Elangovan and Shapiro, 1998). In the same vein, individuals may not risk investing in developing new relationships (e.g., new friends in close relationships) if they already have several ongoing relationships (Hardin, 2000).

Proposition 14 Satisfaction With the Relationship*: In a social exchange, the satisfaction of partners increases with the performance of partners and the perspective of continuity of the relationship, and decreases with the perception of an inequity in the relationship and the existence of several*

others ongoing relationships.

If we add to the satisfaction with the relationship some form of utilitarianism, we are able to consider that the partners to the exchange developed a calculative commitment that will lead to the mutualistic form of benevolence (Lee et al., 2008). In fact, individuals – specially those engaged in economic relationships – are more willing to rely on partners when they expect that the interaction with these partners brings more benefits than costs (Ireland and Webb (2007)).

Proposition 15 Mutualistic Benevolence*: In a social exchange, the satisfaction with the relationship and the exchange compensation associated with the interacting partner increase the calculative commitment of partners, which in turn lead to the mutualistic form of benevolence toward each other.*

Integrity

Cox et al. (2012) provides a thorough philosophical perspective on integrity through the different accounts of integrity that are reported in literature. In a general sense, a person of integrity makes a reflection and self-assessment about her different commitments, wishes, and changing goals, and tries to balance them in order to maintain integrity. She will act on her commitments even when acting on them is hard and she will accept the consequences of her convictions. In order to account for integrity, these commitments must have a social character, imply a proper respect for the deliberation of others, and most of the times relate to moral constraints; at least, they shall be recognized as of great importance by reasonable individuals. In fact, certain persons may act immorally and still consider they are acting with integrity because they may not be aware of their mistaken moral views.

Integrity relates to the general character of individuals. There are a number of traits that prevent an agent to make the change that is needed to act with integrity, such as arrogance, dogmatism, fanaticism, monomania, preciousness, sanctimoniousness, rigidity, capriciousness, wantonness, triviality, disintegration, weakness of will, self deception, self-ignorance, mendacity, hypocrisy and indifference. Some people are more prone to these traits than others. However, as referred by Connelly et al. (2006), the exact boundaries and inner nature of integrity and their relation to individual differences are still unclear.

Scholars on trust tend to define integrity as the trustee's commitment to the principles acceptable by the truster or, more generally, to a set of sound

moral and ethic principles (e.g. Mayer et al., 1995; Elangovan and Shapiro, 1998; Xie and Peng, 2009; Adali et al., 2011). Individuals at higher levels of moral development tend to not trivialize trust violations and are less likely to switch to a different set of principles due to external reasons, thus scoring higher values for the integrity dimension (Elangovan and Shapiro, 1998).

When assessing the integrity of a trustee concerning a given commitment, the truster shall seek for hints about the trustee's capacity to fulfill promises, keep consistency in his actions, be fair, open and reliable, keep value congruence and compliance with social norms (Mayer et al., 1995; Becker, 2005). Any kind of expediency, artificiality, or shallowness shall alert for lack of integrity (Cox et al., 2012). For example, in business, a truster may question the integrity of a firm when its track record with other firms is inconsistent with its stated policies (Schoorman et al., 2007). As another example, a faulty firm that shows sincere apology and regret after a betrayal through acceptance of blame and responsibility, and that implements informational recovery efforts, improve the chance to recover some of its integrity as perceived by its partners (Xie and Peng, 2009). In the same way, the student that failed to return the books to the library on time because he opted instead to watch a movie failed to make a serious attempt to fulfill his commitment and we was not acting with integrity.

However, the assessment of the integrity of a trustee may be affected by the interpretation of the context of the relationship. Mayer et al. (1995) provide the example of the middle manager that makes decisions that appear not to be consistent with previous actions; without further knowledge, employees may question the integrity of the manager. However, if the employees acquire the knowledge that the manager is following orders from top managers, they can stop questioning the manager's integrity.

Following Cox et al. (2012), the fact that an agent acts with integrity in one sphere of his life does not necessarily mean that he is going to act with integrity in other aspects of life, due to the capacity and need for compartmentalization of human beings. For example, an individual may consider that he would never be able to assault a bank, and still he may regularly download files from the Internet without respecting the associated copyrights. Even though, the kind of reflexion of the agent when deciding to remaining true to a specific commitment may flow to other spheres of his life.

Proposition 16 INTEGRITY*: Each individual has a specific disposition to act with integrity – i.e., to remain true to his relevant social commitments – that is related to his traits of personality. However, acting with integrity*

in one sphere of the individual's life does not necessarily assure that he is going to act with integrity in another sphere of life.

Proposition 17 Perception of Integrity*: An individual that is consistent with his actions, shows reliability, value congruence and compliance with social norms is perceived by others as acting with integrity.*

Interrelationship of Ability, Benevolence, and Integrity. Trustworthiness dimensions are generally independent of each other and have unique relationships with trust (Colquitt et al., 2007). Levin et al. (2004) provide an example in the domain of knowledge sharing where an individual is trusted as competent because he knows the information needed by the truster but has not shown any attachment to the truster, showing low benevolence toward him. Another interesting example is given by Mayer et al. (1995): a given trustee is perceived as having high ability and low integrity; however, he shows high benevolence toward the truster. Shall the truster trust him, or shall he fear a betrayal somewhere in the future due to the trustee's low integrity? In fact, it is not trivial to determine the exact balance that needs to exist between the trustworthiness dimensions in order to assure trust, and the perceived lack of any of them may undermine trust. Hypothetically, the truster's propensity to trust will eventually determine the trust he puts on the trustee, taking into consideration the moderating role of this propensity on trust even after trustworthiness information on the trustee is known (Mayer et al., 1995; Colquitt et al., 2007).

On the other hand, the development of the relationship between truster and trustee may also change the relative importance of each of the dimensions of the trustworthiness: integrity and ability data are easier to obtain through third-party sources at the beginning of the relationship, while the impact of benevolence may only be perceived as the relationship grows (Mayer et al., 1995; Schoorman et al., 2007).

Proposition 18 Perception of Individual Trustworthiness Dimensions*: The ability and integrity of an individual may be perceived by a partner earlier in the relationship, through information provided by third-parties. On the contrary, the perception of this individual's benevolence may only be perceived at later stages of the relationship.*

In the same way, the situation may also alter the relative importance of these dimensions (e.g., Mayer et al., 1995; Hardin, 2002; Levin et al., 2004). For instance, in situations where the task in hands does not require complex skills, benevolence and integrity may be viewed as more important than the

competence dimension. As an example, ability seems to have a major impact on knowledge transfers involving highly tacit knowledge, while benevolence is significant in both explicit and tacit knowledge exchanges (Levin et al., 2004).

Proposition 19 Relative Importance of Trustworthiness Dimensions*: The relative importance of the trustworthiness dimensions of trust depends on the task and situation at hands, and on the stage of the relationship existing between truster and trustee.*

Additional Dimensions. Some trust scholars consider different or additional trustworthiness dimensions. For instance, Castelfranchi and Falcone (2010) refer that competence, predictability, and safety are three necessary dimensions of trustworthiness. According to the authors, *predictability* relates not only with the ability of the trustee in doing the task, but also with his *willingness* in doing it. Adali et al. (2011) consider that predictability influences the uncertainty involved in the trust evaluation. Mayer et al. (1995) clearly separate predictability from willingness. The authors argue that a self-interested trustee may be predictable due to the consistency of his past actions, but that he cannot be trusted by a truster if his actions are against the interests of the latter.

This notion of willingness, thus, seem to relate to the consistency, promise fulfillment and reliability attributes of the integrity dimension, and to the sense of commitment that is inherent in the definition of benevolence. Elangovan and Shapiro (1998) refer that all three ability, benevolence, and integrity provide a measure of the trustee's likelihood of keeping (or betraying) the truster's trust. In this sense, a trustee that has a high degree of benevolence and integrity toward the truster in a specific relationship has a low motivation to betray the truster, which translates into the willingness of not exposing him to harm.

In a similar vein, Castelfranchi and Falcone (2010)'s notion of *safety* contribute to the perception of the unharmfulness of the trustee, which is subsumed in the benevolence dimension in Mayer et al. (1995)'s model.

2.3.2 Propensity to Trust

The propensity, or disposition, to trust is a personality trait of the truster that is stable across situations. This propensity is different from truster to truster, due to different personality types, development experiences and cultural backgrounds (Mayer et al., 1995; Schoorman et al., 2007; Tullberg,

2008). For instance, Kiyonari et al. (2006) performed an empirical study using American and Japanese participants and concluded that, in the context of the study, the former were higher in trust than the latter. This definition matches Rotter (1967)'s notion of interpersonal trust. In fact, in his scale for the measurement of impersonal trust, Rotter (1967) measures a kind of generalized trust of others, through generic items such as "Parents usually can be relied upon to keep their promises".

At the extreme case, individuals that repeatedly trust in situations that do not warrant trust to most people may possess what is called *blind trust* (Mayer et al., 1995). This seem to be corroborated by recent studies on neuroscience that suggest that individuals with a certain kind of amygdala damage do not seem to have a normal sense of distrust and danger (Koscik and Tranel, 2011). Conversely, individuals with an increased amygdala activation tend to be associated with social phobia and social avoidance behaviors (Koscik and Tranel, 2011), making us thinking that it would not be terribly exaggerated to extrapolate that these individuals can also be considered as low trusters.

Trust propensity, thus, can be seen as a factor that highly influences the trust that a truster has for a trustee *prior* to data on that trustee is available (Mayer et al., 1995; Colquitt et al., 2007). Some scholars claim that trust propensity is even relevant *after* information about trustworthiness has been gauged, allowing to shape the available trustworthiness information (Govier, 1994, and Lewis and Weigert, 1985, cited by Colquitt et al., 2007). In their study using meta-analytical structural equation modeling, Colquitt et al. (2007) confirmed the relevance of trust propensity in the presence of trustworthiness, although the magnitude of the relationship between trust propensity and trust when trustworthiness was simultaneously considered was relatively weak.

Cvetkovich, Siegrist, Murray, and Tragesser (2002) also agree that trust propensity is able to influence trust after data on the trustee is available – i.e. previous beliefs on trust *persevere* –, although the authors consider trust propensity as a dispositional form of trust that is not necessarily stable across situations. In their study, the authors conclude that individuals may be low or high in general trust of the nuclear power industry and that individuals low in general trust in this specific situation "judged both bad and good news as less positive than those high in general trust" (Cvetkovich et al., 2002).

Hardin (2002) presents a different perspective concerning the influence of disposition on trust, by assuming that trust is little more than knowledge and that the explanation of trusting in some context is "simply an

epistemological, evidentiary matter (...)[and] not a motivational problem". According to the author, trustworthiness, and not trust, can be explained as dependent of motivation, and disposition to trust should not be understood as different from learning how to judge trustworthiness. Based on we have written before, we tend not to agree with such a view.

Although we have focused on trust propensity of individuals, Schoorman et al. (2007) propose that this propensity extends to organizations – developing from geographic, industry, and economic histories –, and that some organizations develop greater propensities to trust than do others.

Proposition 20 PROPENSITY TO TRUST*: Individuals have a propensity to trust that is related with their traits of personality. This propensity influences not only the perception of the trustworthiness of the trustees but also the trust that the truster has on the trustees. This influence is stronger when the truster does not have much information on the trustee. The propensity to trust can extend to organizations.*

2.3.3 Physical and Cultural Characteristics of the Trustee

Moral sentiments, facial expressions, and physical and social characteristics may provide some of the signs that promote cooperation and trust behaviors between strangers (e.g., Hardin, 2001; Kiyonari et al., 2006; Foddy et al., 2009; Kelton et al., 2008; Platek et al., 2009; Castelfranchi and Falcone, 2010; Bzdok et al., 2011; Venanzi et al., 2011). In fact, individuals are able to typecast other individuals, considering certain types of trustees to be more trustworthy than others (Hardin, 2001), probably through processes of primary and secondary socialization (Tullberg, 2008). Moreover, involuntary expressions of moral sentiments and emotional states can also provide clues for trustworthy behavior (Frank, 1988, cited by Allison (1992)).

On the one hand, social and cultural characteristics of agents – such as social category, organizational role, demographic similarity, and cultural relatedness to the truster – can provide clues about these agents' trustworthiness (Allison, 1992; Gambetta, 2000; Levin et al., 2006; Foddy et al., 2009; Hermoso et al., 2009; Adali et al., 2011; Venanzi et al., 2011). For example, the meaning that social norms ascribe to a doctor or a parent brings an associated certification in health caring and parenthood contexts, respectively (Adali et al., 2011; Venanzi et al., 2011). In this case, these social norms provide clues related to the ability of individuals in a given matter and context.

On the other hand, a truster perceives an increased trustworthiness of a stranger that shares with her/him a salient social category if the truster

and the stranger belong to the same group and the truster acknowledges that the stranger is aware of their group membership.[10] This means that trusters have the expectation of altruistic and fair behavior toward fellow in-group members (Foddy et al., 2009). In the same vein, as we have already mentioned in this thesis, recent research on the role of the amygdalae in processing of trustworthiness cues (cf. these studies based on functional imaging and through games, such as Platek et al., 2009) advance neurological-based explanations for pro-social behavior toward kin, closely related with benevolence.

The reviewed literature is not consistent when relating these physical and cultural characteristics directly with trust or with the trustworthiness dimensions, although the latter is more consistently reported. For example, Kelton et al. (2008) propose that this identification with the trustee influences the perception of his trustworthiness, although they do not discriminate amongst the trustworthiness dimensions. Also, it is known that when quick decisions are needed, the facial features of the trustee may have even more impact on trust than specific information about this individual; these facial judgments, which are processed in less than 100 milliseconds, provide pivotal information about the trustee's trustworthiness (Adali et al., 2011; Bzdok et al., 2011). In our previous analysis, we referred the effect of identification on benevolence (cf. Figure 2.2). It also plausible to associate certain physical and categorical characteristics to integrity; in our opinion, it justifies why so many elderly rural Portuguese men and women are deceived by good looking and dressing smart (quack) men, giving them the savings of a life. This is consistent with Dion et al. (1972)'s idea that humans ascribe positive characteristics, such as honesty, to attractive people, even if they do not consciously realize that. In (Venanzi et al., 2011), the authors propose that professional, 'crosscutting' (e.g., male/female) and dispositional (e.g., cautious behavior) categories may influence the perception of the trustee's ability and willingness.

Taking into consideration what was said in this subsection, we formulate two new propositions, that follows, and illustrate the effect of social categorization and relatedness on trust and trustworthiness dimensions in Figure 2.3. Dashed lines represent the ambiguity still existing in trust theory concerning the exact way these social attributions affect trustworthiness and trust.

[10]However, if for any reason this expectation cannot be formed (e.g. when the trusting agent does not have information about the trustee awareness of the in-group situation), the trusting entity should not trust in-group members more strongly than out-group members (Foddy et al., 2009).

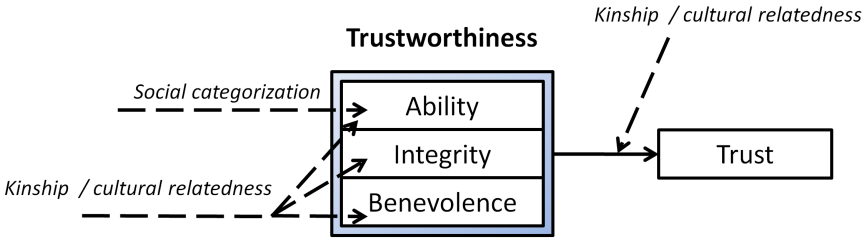

Figure 2.3: The effect of social categorization and relatedness on trust.

Proposition 21 Effect of Social Categorization on Trust*: The social category, organizational role, and other features that socially characterize an individual, such as her/his gender, may provide clues about the individual's trustworthiness in a given matter.*

Proposition 22 Effect of Kinship/Cultural Relatedness on Trust*: Kinship and Cultural Relatedness influence not only the perception of benevolence of the trustee but also the perception of his general trustworthiness. These factors may also influence trust directly.*

2.3.4 Emotional State of the Truster

Emotional states, even when unrelated to the trustee or situation, affect trust (Schoorman et al., 2007). In this respect, Dunn and Schweitzer (2005) refer that incidental emotions with positive valence (e.g., happiness and gratitude) increase trust, and that emotions with negative valence (e.g., anger) decrease trust. Also, the influence of emotions on trust is felt mostly when trusters judge their trust in acquaintances, and less when they judge familiar trustees.

Despite the obvious relevance of the topic of emotions on (computational) trust, we could not explore this research line further in this thesis, due to the complexity of the theme – deserving a thesis by its own – and to time constraints.

Proposition 23 Effect of Emotional State on Trust*: Emotional states, even when unrelated to the trustee or situation, affect trust. This influence is more notorious when judging trust in acquaintances than when judging familiar trustees.*

2.3.5 Reputation

Merriam-Webster defines reputation as the "overall quality or character as seen or judged by people in general" (Merriam-Webster.com). The reputation of a given trustee is the result of the process of social transmission of opinions, general information, *images*, beliefs, meta-beliefs and other social evaluations about the properties of the trustee – his attitudes towards some socially desirable behavior – that circulate over a network of contacts (Conte and Paolucci, 2002; Sabater-Mir et al., 2006; Paolucci and Conte, 2009). Reputation information is by nature more vague than opinions. Most of the times it is about general characteristics of the trustee (e.g., "They say that he is not reliable", or "He is a stone heart"). Other times, however, it mentions specific characteristics or abilities of the target agent, as shown in these examples taken from Merriam-Webster: "He has the reputation of being clever"; "He has earned a reputation as a first-class playwright"; "A teacher with a reputation for patience"; and "Poor customer service has ruined the company's reputation". These cases show a clear connection between reputation and trustworthiness.

Several authors consider reputation as an antecedent to trust (e.g., Sabater, 2003; Patel, 2006; Jøsang et al., 2007; Kelton et al., 2008), in the sense that an individual is more likely to trust a trustee if the latter is trusted by others as well (Kelton et al., 2008). The exact contribution of reputation to trust may depend on the existence and relevance of other types of evidence: "I trust you because of your good reputation" and "I trust you despite your bad reputation" (Jøsang et al., 2007) are both plausible, in this sense. In the same way, if the evidence on the trustee is not enough to make the truster know his trust on the trustee, it is possible that a very high or a very low reputation would allow the truster to mature his mental state about this trust. Thus, reputation may be seen as a "last resort" trust, taking into consideration that epistemological constraints do not allow the possibility that one can trust very large numbers of people through their reputations (Hardin, 2000).

However, one could also argue that knowing the trustee's reputation would not affect the truster's trust on him, but rather any decision that the latter might take concerning being dependent on an action of this trustee. In this alternate view, both trust and reputation are complex and isolated social phenomena, where the process of building reputation is subject to specific social influences that are not present in the process of building trust, such as badmouthing and win-lose games. Moreover, in contrast to what happens in the transmission of trust opinions by honest agents, the agents

that spread the reputation information do not necessarily believe its content, commit to it, or even are responsible for it (Paolucci and Conte, 2009). Hence, in this view, both trust and reputation contribute, in conjunction with other factors, such as risk and utility, to the final desideratum of decision making, and reputation *would not* have a direct influence on trust. This would match the meaning ascribed in certain game-theory and sociological literature where reputation is "... information that agents receive about the behaviour of their partners from third parties and that they use to decide how to behave themselves" (Paolucci and Conte, 2009).

Proposition 24 Reputation*: Reputation is a general and vague information about the quality or character of an individual or collective entity that results from the social transmission of information about this target entity. The information conveyed by reputation is not necessarily true and the people that transmit this information do not necessarily believe its content.*

We think that the exact relation that exists between trust and reputation needs further study from different areas of research, before it can be correctly modeled and implemented by computer scientists. For now, we tend to accept that reputation can act as both an antecedent to trustworthiness and an antecedent to trust, and that its impact on both constructs is rather weak when the truster is secure about his trust. Moreover, we believe that trusting someone through his reputation may occur only rarely. Based on this intuition, we postulate the following proposition.

Proposition 25 Reputation as an Antecedent to Trustworthiness and Trust*: Reputation is an isolated social phenomena that has a moderate impact on trustworthiness and trust, whose strength is residual when the truster is secure about his trust on the trustee.*

2.3.6 Integrative Models of Trust

We overview four conceptual models that try to explain the relation between trust and its antecedents.

Mayer, Davis, and Schoorman (1995)

Mayer et al. (1995) integrative model of organizational (dyadic) trust, illustrated in Figure 2.4, relates trust with its antecedents and consequences; it was designed to be generally applicable and used across multiple disciplines. This model considers two antecedents of trust: the perceptions about the trustee's trustworthiness – namely, his ability, benevolence and integrity –

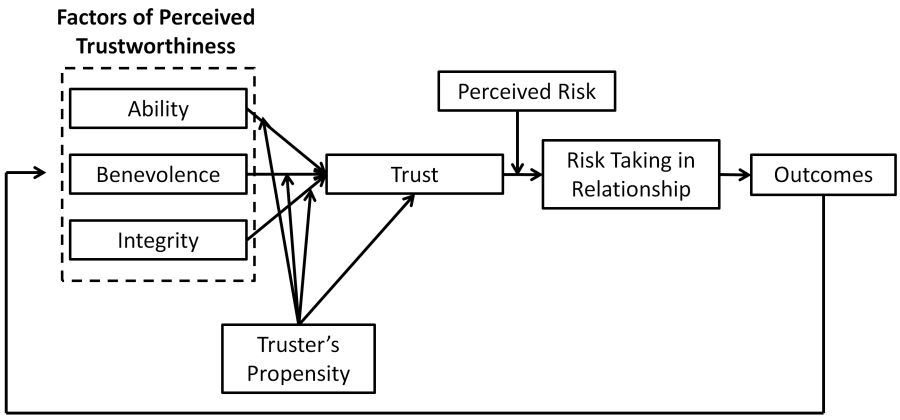

Figure 2.4: The relationship between trust, trustworthiness and propensity to trust, as viewed in Mayer et al. (1995)'s integrative model of organizational trust.

and the truster's propensity to trust, which the authors consider to vary with the personality, developmental experiences and cultural background of trustees. In their view, both antecedents impact how much the truster could trust the trustee, i.e., how much risk he is to take with the trustee. Also, the perceptions of the trustee trustworthiness depend on contextual factors.

The consequences of trust in this model are the actual risk taking in relationship, whose outcome will influence the perceptions of ability, benevolence, and integrity at the next interaction.

Castelfranchi and Falcone (1998)

The socio-cognitive model of trust of Castelfranchi and Falcone includes a comprehensive set of features grounded on the theory of trust (Castelfranchi and Falcone, 1998; Castelfranchi et al., 2003; Castelfranchi and Falcone, 2010; Venanzi et al., 2011). It considers that trust implies the truster to have a given goal that can be accomplished by an action of the trustee, and to believe that he is dependent of the trustee. The trust for the trustee in a particular situation is formed by considering the different beliefs that the former has about the latter, which can be considered as internal attributions (for example, beliefs on competence, disposition, and unharmfulness) or external attributions (opportunities and dangers). These beliefs are fed from four distinct types of belief sources – direct experience, categorization, reasoning, and reputation –, and their values are further modulated by meta-beliefs about the relative strength of each one of them (e.g., how much the source of the belief trusts its own judgment about the trustee,

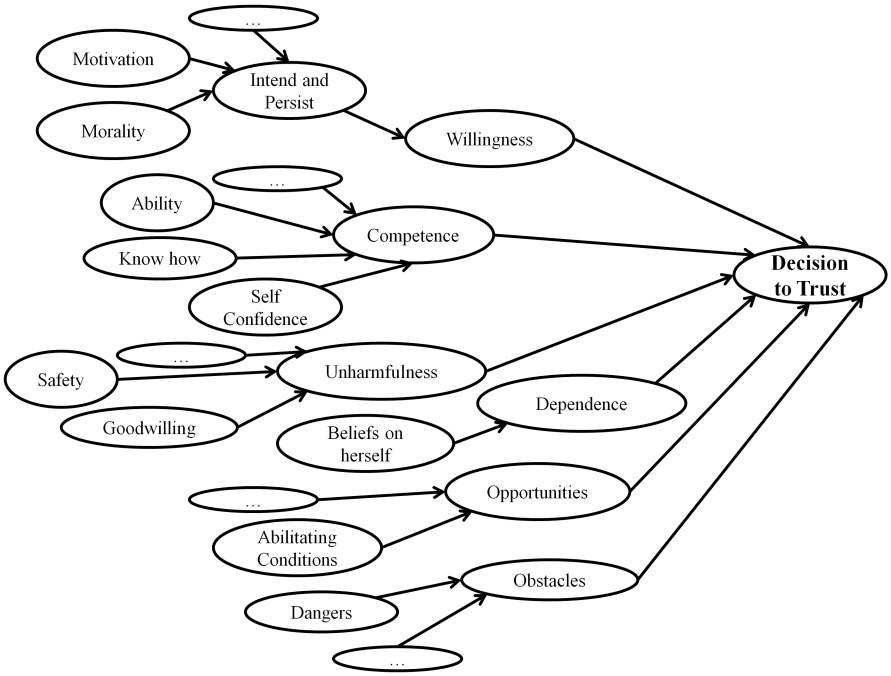

Figure 2.5: The socio-cognitive model of trust (adapted from Castelfranchi and Falcone (2010)).

how much the truster trusts the source, and how much is he certain that the source reported exactly what he understood). The final trust decision conjugates the top beliefs about the trustee's competence and reliability and the relevant contextual factors (see Figure 2.5), where the truster is able to ascribe different weights to each belief taking into consideration the kind of task and his own personality. Hence, the authors consider that trust is a mental attitude toward the trustee and a decision to rely upon him. They propose a three-layer approach to trust which also includes the behavioral account of trust.

In Castelfranchi and Falcone (2010), the authors consider that there is an affective-based form of trust that is not necessarily based on beliefs but instead constitute an emotional reaction – based on automatic and frequently unconscious somatic responses – that appraises the trustee. Sometimes these emotions support the reasoning process, sometimes they are produced by this cognitive process. However, the authors do not model the relationship between trust and emotions.

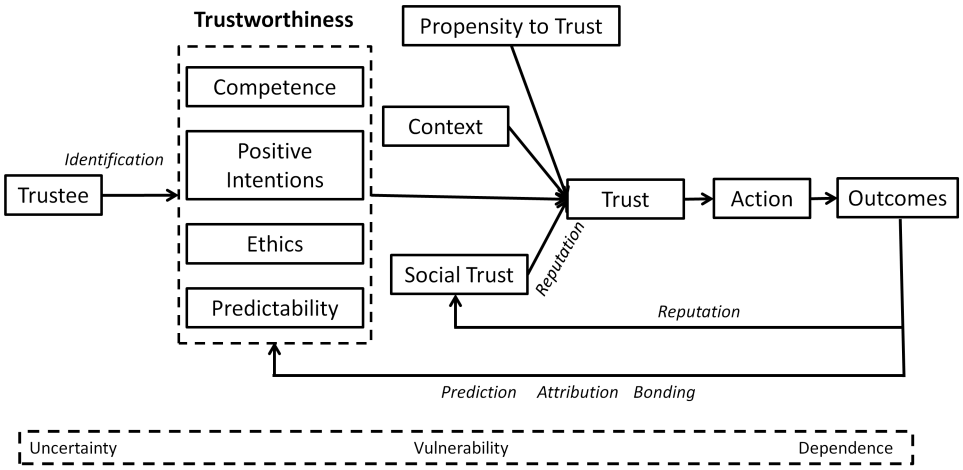

Figure 2.6: Kelton et al. (2008)'s integrated model of trust.

Kelton et al. (2008)

Kelton et al. (2008) presents an extension to the integrative model of Mayer et al. (1995) that includes the preconditions for trust (uncertainty, vulnerability, and dependence), the influence of context and social trust, and the roles of trust development processes. The proposed extensions are the outcome of the authors' theoretical formulations, and the entire model is depicted in Figure 2.6.

The authors consider that uncertainty, vulnerability, and dependence between the truster and the trustee are preconditions for trust. In the same way, they consider that the development of trust is affected by the trustee's trustworthiness, the context, the truster's propensity to trust, and the social evaluation of the trustee by others, transmitted through reputation. In turn, four factors contribute to better perceive the trustworthiness of the trustee: prediction, which relates to the consistency of the past actions of the trustee; attribution of the qualities of the trustee based on observable evidence; bonding, which refers to the development of an emotional relationship between truster and trustee; and identification, which relates to the potential perception by the truster of common identity, goals and values shared with the trustee. Contrary to what happen in the model of Castelfranchi and Falcone, reputation in this model is an antecedent of trust but not of trustworthiness.

Consistently with the models analyzed before, this is a rich conceptual model that lacks validation, although it is grounded on the theory of trust.

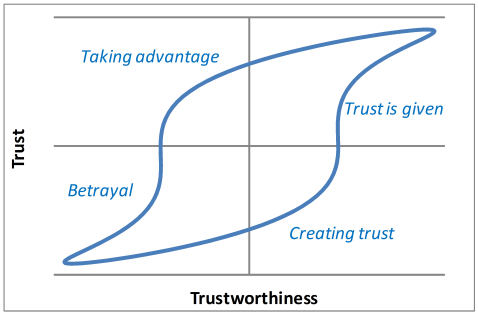

Figure 2.7: The hysteresis of trust and betrayal

Straker (2008)

Straker (2008) presents a very simple model of trust where trustworthiness is the only antecedent of trust. Although the author does not provide hints about the scientific grounds of the model, it presents the interesting characteristic of plotting the dynamics of trust over time, which was identified as important in Mayer et al. (1995). Hence, Straker's principle of the hysteresis of trust and betrayal (illustrated in Figure 2.7) states that when a trustee is not recognized by the truster as trustworthy, he needs to increase his (real) trustworthiness for a long time before the truster can trust him. Eventually, there is a point in this path when he is fully trusted by the truster and he may take advantage of the fact that trust is no more constantly verified. However, if his trustworthiness drops for a long time, the trusting agent will realize that and feel betrayed, with damaging consequences for trust.

Other then the temporal plotting, we think that the hysteresis form ascribed to the dynamics of trust is not grounded. For instance, it not intuitive that a truster, for the same level of trustworthiness, always trust more the trustees that once were fully trustworthy to those that are consolidating their trustworthiness.

2.4 Sources of Trust

In this section, we analyze the sources of information that potentially may provide the truster with "good reasons" and emotional incentives to trust.

2.4.1 Information Sources

Direct Contact. By interacting directly with the trustee, the truster may acquire a credible (probably the best, in his point of view) perception of the

ability, integrity and benevolence of the interacting partner. However, the effectiveness of this source is restricted to the existence of multiple and repeated interactions with the trustee, which is not guaranteed in social and economic environments characterized by high openness and dynamicity. Moreover, interacting directly with a trustee has the cost of knowledge acquisition, which may be specially relevant when the trustee is a stranger and there is a higher possibility of undesirable outcomes.

We include in direct contact the observations made by the truster of the trustee's behavior, even when interacting with others.

Opinions. This is an indirect source of information, where the truster searches his network of direct contacts and asks these contacts for opinions about the trustee, obtained by direct contact. These opinions are subjective and imply that the truster has some mechanism of certifying their credibility and relevance. The opinions can reflect vague information about the trustee – e.g., "I like him" or "I trust him" –, or a more detailed characterization of his trustworthiness, for example: "He can do the task, but sometimes I question his integrity", or "Although he is not very competent, he gave me all he could").

Reputation. In Section 2.3.5, we referred to the two antagonistic views of reputation as an antecedent to trust (and/or trustworthiness) or as an antecedent to decision making. The literature on computational trust seems unanimous in adopting the first view. In this perspective, reputation is commonly seen as a relevant source of information in open and dynamic environments, where other types of information about the trustee can be either inexistent or costly. It is characterized by being highly available, but also by having low credibility, due to the bias introduced by partial reporters and to the noise inherent in multiple transmissions, where rumors and gossip spread easily (Conte and Paolucci, 2002; Paolucci and Conte, 2009; Venanzi et al., 2011).

Trusted Third Parties. Information from trusted third parties, such as certificates (Pavlou et al., 2003) and contracts, provide objective and "safe" information. Availability and affordability are issues to take into consideration.

Categorization, Stereotyping and In-Group. We have mentioned before that faces, cultural relatedness and social/professional categories pro-

vide important perceptions about the trustworthiness of trustees, which motivate us to include them as sources of information.

Emotional States. Strong positive or negative affect for the trustee, as well as the emotional state of the truster, may have a high influence on trust. Lewis and Weigert (1985) refer that the emotional content of most interactions among bureaucrats may be minimal and highly intense when it comes to relations between lovers.

2.4.2 Credibility and Relevance

When using indirect information sources, the truster must check for the credibility and relevance of the reported information. Demolombe (2011) refer that the credibility of information sources can be measured in six different axes: trust in sincerity, trust in competence, trust in vigilance, trust in cooperativity, trust in validity, and trust in completeness. Paglieri and Castelfranchi (2012) consider that source quality must be measured in terms of competence, understanding, and honesty. Besides, they emphasize the role of relevance in trusting information sources: a given piece of information may be true and correctly reported, but is not relevant for the matter of the truster.

Proposition 26 Credibility and Relevance of Trust Sources*: A truster that uses information from others through opinions and reputation shall check the credibility and relevance of the transmitted information.*

2.4.3 Ignorance and Contradictory Information

The gathered evidence and/or mental states provide the grounds for judging the trustworthiness and consequently for trust or distrust. However, Hardin (2004) refers a state of *ignorance* where the truster neither trusts or distrusts the trustee. For instance, the trust one's put on a stranger passing on the street have limited cognitive content (Lewis and Weigert, 1985). At most, particular characteristics of the stranger may trigger some emotional bond or reveal clues of his trustworthiness, as we have already mentioned. Hence, we may argue that the truster is generally ignorant about the trustee's trustworthiness and that he would need more evidence in order to form his mental image of trust.

A distinct concept is trust in the presence of contradictory information. In this case, the truster may have several individuals items of evidence and

still they contain contradictory information, leading to uncertainty. Castelfranchi and Falcone refer that all trust models should include the subjective propensity of the truster to accept a given degree of uncertainty and of ignorance. Also, they consider that the combination of different information sources is a classical complex problem, specially in the presence of diverging beliefs: if someone reports that Mary dresses a hat and some other person that she does not dress a hat, the truster cannot infer that Mary dresses half a hat (Castelfranchi and Falcone, 2010). Should the truster suspend his judgment, take into consideration the best opinion, or the worst opinion? For now, the authors do not present a model to combine beliefs other than the procedure of summing up all the contributions and squashing the result with a threshold function.

Proposition 27 Ignorance and Uncertainty*: A truster that do not possess enough evidence to make a judgment about a trustee are in a state of ignorance. Trusters that have contradictory information about the trustee and do not have any bias from their propensity to trust or mental states are in a state of uncertainty. In both cases, they need additional information in order to make a trusting judgment about the trustee.*

2.5 Trust Dynamics

Social interactions are traditionally secured by ongoing relationships and/or governance mechanisms such as monitoring, contracts, incentives, and institutions. Control mechanisms can have social and economic costs and are not always effective. In opposition, the establishment of long-term relationships is cost effective and is widely used in one-to-one relationships and in commercial relationships. However, the reality of present days indicates the urge for new forms of relationships, mainly in business and in social networks, where relationships are formed more quickly and, more and more, with anonymous others, or strangers. In these new situations, the truster may not be able to ground his trust in the partner through ongoing relationships, because they take time to establish, and the use of institutional back-up may be inadequate. Hence, important questions arise: how individuals trust in the new paradigm of relationships? Can strangers be trusted? Do partners ever evolve into trustful relationships? And how do they react to trust violations in the form of betrayals? In this section, we address the dynamics of trust since its formation to the establishment of ongoing trust relationships. We take special attention to the evolution of trust over time, i.e., to the time dimension referred by Schoorman et al. (2007).

2.5.1 Formation of Trust

People generally do not have incentive to trust when they enter a relationship (Hardin, 2001). In the extreme situation of being totally ignorant about the object of trust, an individual can gamble, but cannot trust (Lewis and Weigert, 1985). Hence, entering an exchange with a stranger may seem surprising at a first sight, specially when there is no prospect of future interaction between the truster and the stranger. However, reality shows that not all strangers are dishonest, and that, in several occasions, individuals that detect some deceitful behavior from their partners are free to walk away from the relationship (Macy and Skvoretz, 1998). Moreover, opportunities for better agreements may exist outside the committed relation (Yamagishi and Yamagishi, 1994).

In Section 2.3.3, we referred that kinship and cultural relatedness may provide clues for trustworthiness that can be used for modulating behavior toward strangers. Levin et al. (2006) refer that the most salient cues available by this time are the trustee's observable features. Other clues for trustworthiness may include reputation, if available, and the use of third-party entities, such as intermediaries that make the social connection between the truster and the unknown trustees (Hardin, 2000). All these can be considered a priori knowledge. This means that a truster may have a perception of the stranger's trustworthiness, but probably he is not confident enough about this estimation in order to form a trust judgment. In this context, Kiyonari et al. (2006) refer that under certain conditions, it is possible to make a trust decision concerning a stranger, but that there is no way that individuals build trust in one-shot encounters.

In the absence of enough information allowing trusting judgments, control mechanisms can be used to protect the relation if those are available and affordable. If the truster decides to interact with the stranger, either by showing a trusting behavior or safeguarded by some other form of control, the outcome of this interaction will allow to form or to update prior perceptions of the trustee's trustworthiness (Mayer et al., 1995). With more evidence about the trustee, the truster can do better decisions about walking away from the relationship, maintaining sporadic interactions with the trustee, or establishing a truly ongoing trust relation with him.

Proposition 28 Trusting Strangers*: Trusters may have a perception of the trustworthiness of strangers provided by observable features of the trustees, reputation and opinions. This perception may or may not allow to form a trust judgment.*

2.5.2 Ongoing Relationships and Reciprocity

Long and stable relations normally provide the conditions and the incentives for trustworthiness and trust (Hardin, 2001). In fact, there are several benefits associated with trust maintenance in a relationship. For example, the partners may enjoy a certain flexibility concerning the fulfillment of contractual obligations and expect that short-term inequities are resolved easily and amicably, or expect improved quality processes (Elangovan and Shapiro, 1998; Schoorman et al., 2007).

Long-term relationships are initiated when one or more parties to the relationship demonstrate benevolence toward the interacting partners. For instance, Ireland and Webb (2007) refer that when a truster faces unanticipated contingencies and yet shows goodwill in detriment of selecting tougher forms of action, he is initiating a norm of reciprocity. At this stage of the relationship ('in-between'), which is neither brand new, nor ongoing, the trustworthiness of partners is based on their trustworthy behavior (Levin et al., 2006).

Proposition 29 Trusting in In-between Relationships*: In in-between relationships, trusters have a perception of the trustworthiness of trustees based essentially on the trustworthy behavior of trustees and opinions. Observable features of trustee and reputation information are decreasing in importance. Trusters may start having notion of a shared perspective with the trustee.*

Eventually, the reciprocation of goodwill actions through repeated exchanges allows the establishment of the norms and shared values that characterize relational behavior, and the perception of trustworthiness is associated with the partners' shared perspective (Levin et al., 2006). Most probably, the relationship will further evolve and *goodwill trust* forms between the interacting partners (Ireland and Webb, 2007; Sako, 2002), where intense emotional investments are usually made; once developed, value human assets and trust are sacrificed with reluctance (Williamson, 1979; Lewis and Weigert, 1985).

Proposition 30 Trusting in Ongoing Relationships*: In ongoing relationships, the perceived trustworthiness of trustee is strongly influenced by the shared perspective of trusters and trustees. Trustworthy behavior and opinions may still influence this perception, but the effect of observable features and reputation is residual.*

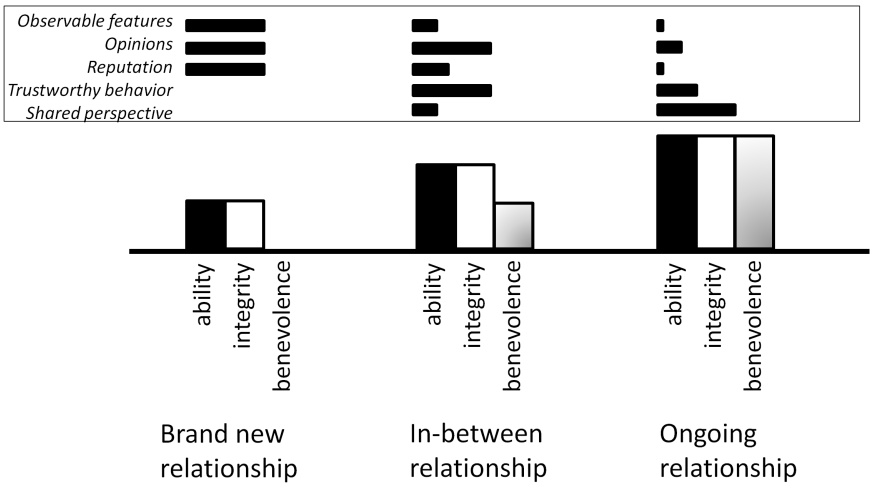

Figure 2.8: Time dimension of trusting relationships: perception of trustworthiness and relevance of individual trustworthiness dimensions.

Reciprocity can then be defined as the mutual exchange of helping behaviors between the partners to the exchange, a kind of tit-for-tat behavior. It might be influenced by several distinct factors; for example, in economic exchanges, we can consider the cultural relatedness, economic nearness, country level risk, governance mechanisms, altruism and reciprocal altruism, exchange of help, and perception of fairness (Lee et al., 2004). In reciprocity-based long-term relations, we can expect that the partners show high levels of benevolence and integrity, in the sense that they commit to the principles accepted by each other. It is also expected that partners do their best to tune hard and soft skills (of all individuals, team, and organization) in order to increase their competence dimension to the level agreed with the interacting partners. Therefore, it is reasonable to think that partners do increase their predicatbility.

This does not mean, however, that trust presupposes reciprocity. The trust invested by a truster in a relationship may not be reciprocated by the partner; in this case, the relationship is broken or never establishes as an ongoing relationship. Figure 2.8 illustrates the importance of the time dimension in relationship and how it influences the perception of trustworthiness and relevance of individual trustworthiness dimensions. Next, we address two characteristics of the evolution of the trust that a truster has in a given trustee over time: asymmetry and perseverance.

2.5.3 Evolution of Trust – Asymmetry and Perseverance

Common sense says that trust is hard to gain and easy to loose. Slovic (1993) studied the thematic in the realm of nuclear power plants, by analyzing the effect that distinct information about positive and negative events had on participants of the study. From the results, he formulated the *asymmetry principle*, stating that negative events tend to have a stronger impact on decreasing trust than positive events on increasing trust.

Other scholars posteriorly conducted similar studies, based on questionnaire surveys, to address the (expressed, or self-reported) impact[11] of positive and negative information on trust. Cvetkovich, Siegrist, Murray, and Tragesser (2002) questioned the implicit assumption made in Slovic's work that people are continually reevaluating and changing their attributions (of trust and distrust), in the sense that established beliefs developed from experience are difficult to change and "exert an influence on the meaning of new information". These authors conducted two studies – one in the nuclear power plants domain and the other in the food domain – from which they concluded that "existing attributions of trust persevere because they affect the interpretation and meanings of new information" (i.e., there is a *confirmatory bias*). Following their conclusions, individuals at a trusting stage tend to maintain or increase trust as they acknowledge positive events, and individuals at a distrusting stage tend to maintain or increase distrust as they learn negative events. Contradictory evidence (positive events when there is distrust / negative events when there is distrust) lead to a discount of information, possibly explained by external factors such as luck/bad luck. In all cases, trust increase is expected to be less than trust decrease, following Slovic's asymmetry principle.

In another similar study concerning genetically modified food in Britain, Poortinga and Pidgeon (2004) reported results confirming that in general terms the asymmetry principle is observed. Also, they verified that participants with clear positive or negative beliefs tend to interpret new information in line with their prior attitudes (the confirmatory bias) but that *ambivalent* participants find information about negative events more informative than negative events (the *negativity bias*). In turn, *indifferent* participants were more unpredictable in their responses and in general terms they seem to suffer the least impact from positive and negative information.

Proposition 31 ASYMMETRY*: The negative events of the trustee tend to have a stronger impact on decreasing trust than positive events on increasing*

[11] In opposition to the *actual* impact on trust (Poortinga and Pidgeon, 2004).

trust.

Proposition 32 PERSEVERANCE*: Individuals at a trusting stage tend to maintain or increase trust as they acknowledge positive events, and individuals at a distrusting stage tend to maintain or increase distrust as they learn negative events. Contradictory evidence may be explained by external factors and do not strongly affect trust at these trusting/distrusting stages.*

2.5.4 Betrayal

We have seen before that trust has a functional role in the continuance of social relationships. However, as put by Lewis and Weigert (1985), this continuance is always problematic, despite the type of the social bond considered. Whatever we think of close relationships, international relations, or organizational relationships, the *breach of trust* is a possibility at any point of the bond, most probably caused by the failure of trustworthiness (Hardin, 2002).

Finkel, Rusbult, Kumashiro, and Hannon (2002) define betrayal in the context of close (established and trustful) relationships as the "perceived violation by a partner of an implicit or explicit relationship-relevant norm". Fitness (2001) extends this definition to "any kind of relationship context" inasmuch as salient relational expectations are violated by any one of the partners in the relationship.

Betrayal is distinguished from other negative incidents in the sense that it involves the violation of the rules (either relationship specific or culturally shared) that govern interaction (Finkel et al., 2002). Elangovan and Shapiro (1998) add that betrayal implies the *voluntarily* violation of *mutual known* pivotal expectations of the victim, potentially causing harm to the latter. They also consider that betrayal is a violation of personal trust and that deviance is a violation of impersonal trust.

The perception of the severity of the violation of pivotal expectations, and the harm it can cause to the victim, may depend on the traits (Fitness, 2001) or even the neurological situation (Koscik and Tranel, 2011) of the victim, in the sense that some actors may feel betrayed in situations where others do not. However, if the victim perceives a betrayal, its consequences are devastating (Fitness, 2001; Cvetkovich et al., 2002; Poortinga and Pidgeon, 2004), because it normally disrupts ongoing and meaningful relationships in which partners have invested material and intense emotional resources (Lewis and Weigert, 1985; Fitness, 2001).[12] In some cases, a single

[12]In her studies, Fitness (2001) proposed that "laypeople hold elaborate theories about the nature of forgivable and unforgivable offenses in marriage".

act of betrayal can destroy trust instantly (Poortinga and Pidgeon, 2004). As interestingly put by Cvetkovich et al. (2002), "In Dante's *Inferno*, the second level of Hell is reserved for those who betray the trust placed in them".

Proposition 33 Trust After Betrayal*: The perception of a betrayal is devastating and destroys the truster's trust on the trustee.*

Assessing the Relationship. Elangovan and Shapiro (1998) present a general model of opportunistic betrayal in organizations that, in our opinion, can be easily accommodated to most generic types of social interactions. In their model, there are certain conditions, such as the presence of a financial crisis or unfulfilled needs of the agent, that trigger him to assess the situation of the relationship. Betraying is one possible outcome of such an assessment when the agent realizes that he is unsatisfied with the current situation; other possible outcomes are for the agent to continue or to abandon the relationship.

The assessment of the situation takes into consideration the benefits associated with betraying the truster versus maintaining the status quo, the relationship with the partner, and the principles (ethics) involved in the decision to betray (or not). Moreover, the assessment of the relationship with the partner is influenced by the perceived equity of exchange, the perceived continuity of the relationship, and the availability of alternate partners. Individuals tend to balance the inputs and the outputs of their relationships. As we have mentioned before, the perception of any inequity downgrades the value of the relationship. However, the reciprocation of inputs into the relationship increases the satisfaction with the relationship (see also Lee et al., 2008). Also, when an individual perceives that the relationship is coming to a natural end, the partner becomes less central in his life. Whether or not this diminishes the satisfaction of the relationship has to be weighted with any existing personal friendship. If the relation is expected to last, the individual tend to deal with the partner's expectations with more care. Finally, if the present situation is ranked poorly, the trustee is motivated to betray. However, the actual decision to betray is influenced by the trustee's perceived likelihood of suffering severe penalties due to betrayal.

A related question is what happens to benevolence after a betrayal. Following Ireland and Webb (2007), the benevolent behavior of a party in the presence of an unexpected contingency depends on the level of trust existing between partners and on the magnitude of the contingency. When an individual feels that he was betrayed, if something can be done by the

perpetrator of the betrayal to rebuild the partner's benevolence, it includes apologetic actions of repairing trust, such as sincere apology (acceptance of blame and responsibility), regret after a betrayal, and respect (Xie and Peng, 2009).

Trust Repair. The first emotional reactions to betrayal in neurological normal adults are pain, sadness and hurt, and it will take some time until the betrayed interprets the situation and eventually feels different other types of emotions (Fitness, 2001; Finkel et al., 2002). Negative cognitive patterns (e.g. confusion, obsession in reviewing prebetrayal events), as well as negative behavioral tendencies toward the betrayer, such as vengeance or demanding of retribution, may also be developed (Finkel et al., 2002).

If the relation can ever be repaired, it will in most cases imply that the victim surpasses a process of *forgiveness* (i.e. giving up destructive behaviors) toward the betrayer (Fitness, 2001; Finkel et al., 2002; Xie and Peng, 2009). Forgiveness will depend on several factors, such as the severity of the betrayal – different kinds of relationships involve different kinds of rules and expectations (Fitness, 2001) –, the emotions and cognitions that accompany the act, the personal values and long-term goals of the victim of betrayal, and the relationship with the transgressor, including her/his perceived commitment (McCullough and Hoyt, 2002; Finkel et al., 2002).

In organizational marketing, Xie and Peng (2009) propose that a firm should use affective, functional, and informational repair initiatives as trust-repairing efforts after negative publicity, in order to demonstrate its integrity, benevolence, and competence during the handling of the crisis.

Affective efforts include an apology toward the victim(s) of the betrayal. If it is based on the sincere repentance of the offender, apology helps reallocating esteem (Xie and Peng, 2009) and redressing the power imbalance between perpetrator and victim (Fitness, 2001). In Xie and Peng (2009), affective recovery efforts improved perceptions of integrity and benevolence. Functional efforts include financial compensation and taking actions to avoid similar violations in the future. The results of the study have shown that these efforts helped to moderately increase the perceived competence of the offender, but had no effect on the perceived integrity or benevolence (Xie and Peng, 2009). Informational efforts include "demonstrating evidence, clarifying facts, and disclosing update news during the crisis handling process" (Xie and Peng, 2009).

The study also showed that forgiveness fully mediated the connection between benevolence and overall trust, whereas it served as a partial mediator in the connections between integrity/competence and overall trust.

Moreover, companies that demonstrated a high capability in crisis handling were able to regain trust directly, without passing through the forgiveness process.

2.5.5 Promoting Trust

In practically all kind of social relationships, the best way to create trust is to be trustworthy. An agent that intends to gain the trust of others shall provide the means that allow the others to acknowledge his trustworthiness. Probably the most effective way to do this is to act in a benevolent way. For example, in economic exchanges or joint ventures, partners may increase the communication, engage in open exchange of ideas, including the exchange of information above normal levels, or even allowing benchmarking in order to increase their trustworthiness, as perceived by the interacting partners (Elangovan and Shapiro, 1998; Schoorman et al., 2007; Xie and Peng, 2009). The perception of the agent's trustworthiness may also increase if his trustworthy actions are indirectly acknowledged by others through reputation. In the same way, in scenarios where relationships are backed up through the reliance on societal and institutional devices, and therefore individuals more easily risk new relations, acting in trustworthy way helps creating one's image of trustworthiness (Hardin, 2002). If the agent is rebuilding his trustworthiness after perpetrating a betrayal or other harmful action, he needs to engage in trust repairing actions in order to elicit the forgiveness of the offended partner (e.g. Fitness, 2001; Finkel et al., 2002; Schoorman et al., 2007; Xie and Peng, 2009).

2.5.6 Promoting the Trustworthiness of Trustees

Sometimes, individuals need to interact with trustees with whom they have little information about their trustworthiness. Hence, they may try to promote the trustworthiness of these trustees, trying to create the grounds for trust by giving trustees the incentive to be trustworthy Hardin (2004). One evident action of trusters is to engage in trusting actions with the trustees. However, it is not evident that trust begets trustworthiness. Kiyonari et al. (2006) analyzed different empirical studies in social psychology in order to understand if trust begets trustworthiness in one-shot encounters. These studies, including the one performed by the authors, were based in trust games, and yield contradictory findings, that the authors linked to the domain specificity, configuration and methodologies of the experimental set.

The truster may also use the law of contracts to help individuals be trustworthy (Hardin, 2004). In the same vein, the partners seeking benefits

in the long run have incentive to be trustworthy (Hardin, 2001).

2.6 Trust and Social Control

Social exchanges involve uncertainty, due to contradictory or asymmetrical information, and risk. Trust involves, *at least*, the willingness to assume risk, while behavioral trust explicitly assumes risk (Mayer et al., 1995). [13]

In order to couple with the uncertainty, risk, and associated vulnerability of partners in everyday relationships, some form of social control is needed.

There is an interesting link between trust and control. Trust indicates the amount of risk that one is willing to take (Mayer et al., 1995; Schoorman et al., 2007); in certain situations, agents may risk the interaction even if the initial level of trust is low (Burnett, 2011). However, most of the time, when trust is not enough, control constitutes an alternate mechanism that allows lowering the perceived risk to a level manageable by trust (Schoorman et al., 2007; Castelfranchi and Falcone, 2010). Control can be defined as the actions that an agent performs or delegates to a third party in order to affect the behavior of others, increasing the agent's belief that the others will behave as expected (Das and Teng, 1998; Burnett, 2011). This is usually accomplished through monitoring and intervention.

For instance, norms constitute shared expectations for behavior, i.e., they are rules universally adopted within a group, which are maintained by sanctions (Allison, 1992). Some of them are legal norms, some others are implicit rules of behavior without legalistic basis. The existence of legal norms is one of the most effective remedies to confine the risk associated with lack of trust, supporting the decision to invest trust in a relationship. Legal regulations and sanctions reduce the risk of being betrayed, by exerting pressure on individuals to conform (Luhmann, 1979; Das and Teng, 1998; Bachmann, 2001). In some conditions, they foster the constitution of trust when it does not exist (Bachmann, 2001). However, when control is too strong, it may have the opposite behavior of inhibiting the development of trust (Schoorman et al., 2007). Also, control mechanisms such as legalistic remedies are usually costly and they are not always effective, but there are situations where the risk of loss justifies the expense of using them (Hardin, 2001).

Other governance mechanisms that may be used to compensate low trust

[13] The relation of trust and normative control is being addressed at LIACC in the scope of project FCT/PTDC/EIA-EIA/104420/2008. Some text in this section might be the result of collaborative work of the PhD proponent and the team project colleagues, and have been adapted from (Urbano et al., 2011a).

include incentives, careful selection processes and socializing efforts (Wathne and Heide, 2000).

2.6.1 Trust in Normative and Contract Systems

Trust is of paramount importance in business. In fact, one of the main factors that inhibits a wider and faster adherence to new technologies, such as electronic payment over the Internet, is (the lack of) trust. A similar dependence is observed when selecting business partners. Trust issues become more important when the acquaintance of potential partners is lower.

Looking precisely at how business relationships are established, Tan and Thoen (2000) proposed looking at transaction trust as composed of two parcels: party trust and control trust. Party trust refers to the trust one has on the other party of a potential business relationship. When such trust is not enough, control mechanisms (e.g. contracts, institutions) must be used in order to enable a business transaction to take place. Control trust refers to the fact that one must trust the control mechanism that is being used. Looking at these parcels as complementary, an agent will engage in the transaction when the level of transaction trust is above his personal threshold. This threshold is determined by the potential gain and risk of the transaction. Considering risk, for instance, the threshold may depend on the type of transaction (e.g. the higher the transaction value is, the higher is the threshold) and on the other parties involved (e.g. the threshold may be lower if the interactions partners are perceived trustworthy).

A typical case of a trading practice in international trade is the letter of credit control procedure (Boella et al., 2005). In this case, the lack of trust between a client and a supplier is replaced by a professional banking relationship between each party's bank. This relationship is more trustworthy because documentary credit procedures are subject to guidelines issued by international bodies, such as the International Chamber of Commerce, and because trading banks tend to have long term relationships supported by good reputation records.

Control mechanisms are used because agents have the expectation that they will somehow make the other party's behavior more predictable. As such, when drafting a contract tailored to a particular transaction and business partner, it turns out that an appropriate enforcement institution must be in place so that the agent can trust the contract contents as a control mechanism.

A few researchers have devoted their efforts on studying the interplay between trust and normative multi-agent systems. Boella et al. (2005) consider

the effect of norm violations on trust. In this perspective, trust amounts to an expectation of the truster towards the trustee of compliance with, for instance, an obligation. Given a violation, trust may or may not be affected depending on the causes for the violation. Boella and van der Torre consider also the internal motivations of agents when fulfilling or violating norms, and the effect of sanctions on their behavior. An agent can be trusted in a specific interaction as long as the sanction is effective in discouraging a potential violation. On the other hand, it may be the case that an agent fulfills an obligation not because of fear of sanction but simply because the agent is respectful. In such a case, the agent could be trusted regardless of there being a sanction.

Employing trust in contracting processes seems to be a natural thing to do. Negotiation and, moreover, contract drafting are typically informed by the trust one has on the potential partners we are dealing with. Even so, the use of computational trust in such processes is not yet much explored by the research community. König et al. (2008) provide a theoretical analysis of the potential use of reputation information in electronic negotiations. These authors study which role(s) participating in a negotiation (taking place at an auction, brokered market or direct bargaining) is in a position to exploit reputation information.

The notion of *sanction* can be analyzed from a broader perspective. An institution may, broadly speaking, apply two basic kinds of sanctions in order to incentive norm compliance; or, to put it another way, to discourage deviations (Pasquier et al., 2005; Grossi et al., 2007). *Direct material sanctions* have an immediate effect, and consist of affecting the resources an agent has (e.g. by applying fines). *Indirect social sanctions*, such as changing an agent's reputation, may have an effect that extends through time. Depending on the domain and on the set of agents that are being addressed, the effectiveness of such sanctions may be different: if agents are not able to take advantage of other agents' reputation information, material sanctions should be used instead.

Some researchers study the use of trust and reputation as non-costly replacements for material sanction-based approaches. For instance, Villatoro et al. (2010) study different kinds of sanctions, both positive (rewards) and negative, that may be used as reinforcement mechanisms that strengthen the fulfillment of norms. However, the authors focus their attention on the so-called *interactionist view*, where norms are seen from a bottom-up perspective, instead of being used as regulatory instruments to govern a specific collective activity (the *legalistic view*) (Boella et al., 2008). A mixed approach seems to be adequate to the domain of B2B contracting: norms

do govern a contractual activity, and as such may impose sanctions in case of non-compliance. Agents may, however, negotiate their contractual norms at runtime, together with associated sanctions.

Despite these above mentioned, mostly theoretical, studies on this issue, the effective and automated use of computational trust and reputation in designing norms (as applied to contracts), selecting control mechanisms and enforcement policies is still lacking. And yet, trust issues seem to be ubiquitous in business relationships, which makes research in this domain very pertinent.

2.6.2 Trust and Opportunism in Business Relationships

Taking into consideration the initial desideratum of developing a computational trust system that could be applied to business relationships, we will narrow the discussion of trust and control to this type of relationships. Most content of trust and social control applies to business relationships, as the nature of these relationships is either economic and social (Lee et al., 2004).

The literature on organizational trust is fertile in the study of uncertainty and opportunism in business exchanges. Usually in this field, uncertainty is harder to reduce through personal relations, and the vulnerability of partners is also harder to reduce due to the presence of power relations between partners (Heimer, 2001). Both uncertainty and vulnerability leads to opportunism, which can be defined as "some form of cheating or undersupply relative to an implicit or explicit contract" (Wathne and Heide, 2000).

Opportunism can be passive or active and applies under existent conditions – evasion of obligations (passive) and violation (active) – or new conditions – refusal to adapt to new circumstances (passive) and forced renegotiation (active); it is either present in informal agreements or legal contracts (Wathne and Heide, 2000). A common problem in interfirm relationships is adverse selection, where suppliers hide their true attributes from the buyer. This happened in the famous Ford vs. Lear case, where Lear committed to supply the seats for all Ford Taurus versions, withholding the information about its lack of adequate resources. As a result, "Lear missed deadlines, failed to meet weight and price objectives, and furnished parts that did not work (Walton 1997)" and Ford incurred in substantial transaction costs (Wathne and Heide, 2000).

Different governance mechanisms are proposed to manage opportunism, from specific forms of control and monitoring to sophisticated selection mechanisms, which can include certification and reputation (Das and Teng, 1998). One of these mechanisms is to use legalistic remedies, including the

use of formal contracts (Sako, 1998). However, designing detailed contracts may involve substantial drafting and monitoring costs (e.g. Williamson, 1979; Macy and Skvoretz, 1998; Gambetta, 2000; Wathne and Heide, 2000; Cvetkovich et al., 2002; Ireland and Webb, 2007), especially when monitoring is difficult or when sanctions require extensive litigation (Macy and Skvoretz, 1998). For example, there are still open issues concerning jurisdiction in e-commerce contracts and the cost of legal procedures may often be higher than the value of contracts (Jøsang and Ismail, 2002). Relational contracts, very used in industry (e.g., in textile industry, Tokatli, 2007), lighten up the drafting of contracts but still have the costs associated to litigation. In general terms, although contracts may bring organizational legitimacy, they are often ineffective (Mayer et al., 1995; Gambetta, 2000), because the focus is shifted from trust to the efficacy of sanctions and to the ability to enforce them when the contract is broken (Gambetta, 2000).

Hence, trust seems a more effective mechanism of social control than sanction-based mechanisms. This does not mean, however, that the latter are not necessary. On the one hand, not all relations are trust relations. On the other hand, all trust relations are subject to endgame effects, and one is better off if secured against these effects, especially when dealing on matters of great importance (Hardin, 2001). Das and Teng (1998) refer that confidence in a transaction may be obtained as a combination of trust and control: for the same level of confidence, if we trust less, we use more control mechanisms. Furthermore, trust and control are seen as parallel and supplementary notions: they contribute independently to the level of confidence, and any one of these mechanisms may be used if an increase in transaction confidence is needed. These two governance mechanisms are therefore interconnected.

2.7 Concluding Remarks

In this chapter, we reviewed important concepts about trust that have been studied over the last decades in several research disciplines. Although there is still a lot of divergence and ambiguity associated to trust theory, there are also mature concepts that can and shall be modeled in computational approaches to trust. Probably the most important concept is that trust is social: it is based on the relationship existing between trusters and trustees, and must account for the time dimension and situation of relationships. Recent work on the theory of trust grounds this concept. In the same way, trustworthiness is just one, though important, antecedent of trust. When estimating a trustee's trustworthiness, it is fundamental to distinguish be-

tween ability, integrity, and benevolence. It is this last dimension that accounts for most of the social nature of trust: if a truster and a trustee are engaged in a dyadic relationship and the trustee acts benevolently toward the truster, then most probably the trustee is going to show his trustfulness to the truster in the future. Other antecedents of trust to be considered are the propensity to trust and the emotional state of trusters. However, as we are going to see in the next chapter, only a few computational trust models address the social account of trust and its situational dependence, and probably even fewer distinguishes between the trustworthiness dimensions of trustees or even consider the propensity to trust of trusters. The influence of the emotional state of trusters on trust is a subject still nebulous even in the theoretical field.

We devoted a great amount of time to the study of social trust, because somewhere in the studying of others' computational approaches and in the development of our own computational approach, we felt that something was lacking – some comprehension of what trust really was. The exception is for the work of Castelfranchi and Falcone on socio-cognitive trust, which presents a thoughtful and detailed study on trust and propose a computational implementation of the model (e.g., Castelfranchi and Falcone, 2010). However, our approach differentiates from the work of these authors in different aspects. First, as we are going to see, we try to give greater emphasis to the relational account of trust. This will reflect more clearly when we present our computational model later on this thesis, where the interpretation of the evidence about a given trustee must reflect the relationships that were established between the trustee and each one of the evaluators of the trustee's behavior. Second, we give particular emphasis to the personality traits associated to benevolence and integrity. In the same way, we make a distinction between trust and acting on trust following Mayer et al. (1995)'s perspective.

For reasons of systematization, and also with the hope that this work might be of interest to other researchers in computational trust (especially those that are giving the first steps in the thematic), we derived a set of propositions capturing the key concepts on trust. In the next chapter, we will overview existent models of computational trust against these propositions. This does not mean that the propositions cover all trust theory or even that computational models shall implement all the derived propositions, but they will help to understand how close or how far the models are from trust theory. These propositions also guided the development of SOLUM, our computational trust model, that we present in Chapter 4.

There are some topics that we did not cover with detail in this chapter,

such as the influence of emotional states on trust and the relation between trust and risk, and others that we have not covered at all, such as the application of trust across levels of analysis (cf. Schoorman et al., 2007), the cross-cultural and gender differences, and the (situational) transitivity of trust. We reserve these topic to future work.

Chapter 3

Computational Trust

3.1 Introduction

Computational trust is a relative new field of research that emerged from the research on different areas – such as distributed problem solving, dependence relations and cooperation among autonomous agents, emerging behavior in agent-based systems, and artificial ethics and morality – in the early 90's of last century. One of the first works that used in some way the notion of computational trust was described in Carley et al. (1993), where the authors simulated different organizational structures that resulted from the combination of three social characteristics of social agents: honesty (vs. lying), cooperativity (vs. selfishness), and benevolence (vs. non benevolence), and where social agents were able to judge the other agents concerning their reliability providing specific information. However, probably the first work that addressed computational trust in a systematic way was presented in Marsh (1994).

Since then, several other computational approaches have being proposed that address trust and its relation with reputation. We do not intend to cover all these approaches here, as some of them are systematically covered elsewhere (see, for instance, Jøsang et al., 2007; Pinyol and Sabater-Mir, 2011). Instead, we propose to analyze how each of the key concepts of social trust that we have captured in the form of propositions in the previous chapter is addressed in different computational trust approaches. As we are going to see next, most of these approaches are more focused on one or two individual trust concepts and fail to address the other concepts. Furthermore, most of the existent computational trust approaches fail to capture the social nature of trust. Finally, we give special emphasis to those works that we think are more representative of the social features in analysis.

3.2 Computational Trust Models

Computational trust models consist of one or more computational components that, working together, take as input the trust evidence available on the trustee under evaluation and output an estimation of the truster's trust on this trustee. This process implies that the computational model must have, at least, the following functions:

1. A proper representation of trust evidence, including the representation of the context of the events from which the evidence is generated, the attributes considered to represent the evidence, and the set of possible values allowed for each attribute of the evidence.

2. A means for acquiring, collecting, and/or generating the individual items of trust evidence; this may include additional processes for inquiring third-party information sources about the behavior of the trustee and inferring the credibility of these sources, which is a field of investigation in computational reputation systems.

3. An aggregation function responsible for generating the trust score from the set of evidence. The resulting trust score is generally included in the decision process of agents concerning the possibility of interacting with the trustee.

In reality, computational trust models that cover, but not expand, this set of functionalities lack, at least, the emotional content of trust; they are mostly cognitive-based processors that estimate the trustworthiness of agents from the reports about their past behavior. Although the emotional content of trust is undeniable (Proposition 2), to the best of our knowledge, there is not any computational trust approach that addresses this in a systematic way. In the same way, the aggregation functions of most computational trust approaches tend to oversimplify the cognitive process that ultimately leads to the estimation of the trustee's trustworthiness score. However, the combination of information from different information sources is too complex when modeling human behaviors, and humans use different heuristics to combine opposite beliefs, as illustrated in Castelfranchi et al. (2003)'s example: "if someone says that Mary dresses a hat and another one says that she does not dress a hat, I cannot infer that Mary dresses half an hat" (although she could be wearing something *close* to a hat!). In the presence of divergent opinions, the trust decision should be guided by criteria linked to context, emotions and personality factors, where some trusters may decide to suspend the judgment, others can consider the best opinion,

others the worse opinion, and so on (Castelfranchi et al., 2003). Hence, a computational trust model shall be able to implement different heuristics, and it must be more than mere statistics or reinforcement learning (Castelfranchi et al., 2003; Castelfranchi and Falcone, 2010). We start this section by briefly mentioning some of the earliest computational trust models.[1] Although they are very simple, they address important features that remain actual.

3.3 Simple Trustworthiness Estimators

The Model of Marsh (1994)

Marsh (1994) presents a model for computational trust that includes the definition of situational trust. Concretely, this model considers that the trust that a truster x has on a trustee y in a given situation α is given by the probability that x acts to achieve an outcome as if he trusts y, $\widehat{T_x(y)}$, weighted by the product of the amount of utility x gains from situation α, $U_x(\alpha)$, and the subjective importance of the situation, $I_x(\alpha)$. Hence, this model mixes the concept of trust with the utility of being in a given situation. In turn, $\widehat{T_x(y)}$ is the general trust of x that is obtained from his trust on y in different situations A, such that $\widehat{T_x(y)} = 1/|A| \times \sum_{\alpha \in A} T_x(y)$.

The SPORAS Model

The SPORAS model (Zacharia and Maes, 2000) uses an update function to estimate the reputation score of a trustee after a new rating on this trustee is received. This function takes as input the trustee's most recent reputation, the reputation of the user giving the rating, and the value of the current rating. Also, using this function, users with very high reputation values experience much smaller rating changes after each update than less reputed ones. In the same way, the most recent ratings have more weight in the evaluation of a particular users's reputation, and unknown agents have always worse reputation than already classified agents. This model also defines a measure of the reliability of the reputation score that is based on the deviation of the estimated reputations.

[1] Some of these models are categorized by their authors as computational reputation systems. However, they present aggregation functions that are used to compute the trustworthiness of agents, and therefore they present no distinction, at the aggregation level, from computational trust models.

The AFRAS Model

The AFRAS (A Fuzzy Reputation Agent System) model (Carbo et al., 2003) is a distributed approach to computational reputation. This model considers that both the reputation scores of trustees and the individual opinions about trustees are represented using fuzzy sets, in order to allow for more natural classification expressions (e.g., 'extremely good'). Every time there is a new opinion, the model aggregates the fuzzy set representing the current reputation score with the fuzzy set representing the new opinion using a weighted means, where the weight determines how much of the previous experience is taken into account, i.e., the *memory* or remembrance of the reputation function. In turn, this weight evolves with the accuracy of previous predictions: the more accurate these predictions are, the more past experience is remembered, and vice-versa.

The Regret Model

The Regret model (Sabater and Sierra, 2001; Sabater, 2003) aggregates impressions to compute the reputation score of trustees. An impression registers an interaction between two agents from the point of view of the truster, relative to some aspect of the interaction. The truster (agent a) keeps a database of impressions about the trustee (agent b), from which it can estimate the reputation derived from direct experience ($R_{a \to b}$), the reputation derived from direct experiences of the group of agents A where a is inserted ($R_{A \to b}$), the reputation derived from the experiences of the trustee's group B with b ($R_{B \to b}$), and the reputation derived from interactions between a's group and b's group ($R_{A \to B}$). Each one of these values of reputation is computed using a weighted mean of the impressions' rating factors, where the weight is a function of the recency of the impression, giving more relevance to recent impressions; the aggregation of these different values of reputation into the reputation score in one aspect of the interaction is calculated using again a weighted mean, whose weights are chosen according to the specific domain under assessment. Finally, the final reputation score of the trustee in one given scenario is given by the aggregation of the reputations calculated for the different aspects that constitute this scenario, which are defined using (ontological based) graph structures. Figure 3.1 illustrates an example of an ontological structure for a good seller, adapted from (Sabater and Sierra, 2001).

Regret also defines a value for the credibility of the reputations score mentioned above, given by the number of impressions used to calculate the reputation value and the variability of their rating values. This variabil-

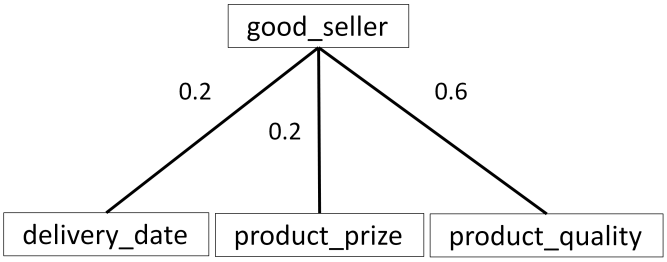

Figure 3.1: An example of an ontological structure for a good seller.

ity, which is similar to the one proposed in SPORAS (Zacharia and Maes, 2000), measures the volatility of the trustee in fulfilling its agreements. It is, then, related to the integrity dimension of the trustee's trustworthiness (cf. Proposition 17).[2] Reputation scores are transmitted through witnesses as a tuple of real values representing the reputation value and the credibility of this value as assessed by the witness. A truster that receives reputation scores from his social relations compute the final aggregated value by weighting each score with the credibility of the witness that sent it.

The FIRE Model

The FIRE model (Huynh et al., 2006) proposes to integrate different sources of trust – direct experiences between truster and trustee, witness reports, third-party references, and rules provided by end users encoding beliefs or knowledge about the environment – in order to provide a collective trust measure. Each one of these sources feeds a different trust function. Reports that result from direct experiences and witnesses' reports are aggregated in a similar way as the average of the correspondent ratings weighted by their recency.

FIRE also proposes to address a specific account of social trust, more specifically, the trust resulting from the role-based relationships between two agents. For this, each agent has a set of domain-specific rules defined by its owner that encode norms of the environment. For instance, the rule `(_, team-mate, honesty, 1.0, 1.0)` tells the agent to expect total honesty from his team mate, and the rule `(buyer, seller, quality, 0.3,-0.2)` tells the agent that ordinary sellers usually sell a product of slightly lower quality than agreed, and that the reliability of this belief is low (0.3). How-

[2] As we will see in Chapter 4, we propose a related measure to infer the integrity of the trustee. However, we integrate the consistency of the trustee's outcomes in the trustworthiness score itself and not in a credibility score of the trust (reputation) score as it is done in SPORAS and Regret.

ever, these rules are hard-coded by the agent's designer, lacking the desired flexibility in dynamic multi-agent systems.

The Beta Reputation System

Jøsang and Ismail (2002) propose the Beta Reputation System, which uses Beta probability density functions to combine feedback and to derive reputation ratings. This model defines the reputation function of target agent T by evaluator agent X, $\varphi(p|r_T^X, s_T^X)$, which is a Beta density function parametrized by r_T^X and s_T^X; the first of these parameters represents the collective amount of positive feedback about T provided by X, and the second represents a similar amount, now concerning negative feedback.[3] Furthermore, it defines the reputation rating of T by X as the probability expectation value of the reputation function:

$$E(\varphi(p|r_T^X, s_T^X)) = \frac{r_T^X + 1}{r_T^X + s_T^X + 2} \ . \tag{3.1}$$

This reputation rating has values in $[0, 1]$, where 0.5 represents a neutral value. When estimating the trustworthiness of T, the reputation system combines feedback from multiple sources by simply accumulating all the received r_T^X and s_T^X parameters from the feedback providers.

Assuming that a collection of agents have provided a sequence Q containing n feedback tuples $(r_{T,i}^Q, s_{T,i}^Q)$ indexed by i about target agent T, the combined reputation function of T in Q, $\varphi(p|r_T^Q, s_T^Q)$, is given by making:

$$r_T^Q = \sum_{i=1}^{n} r_{T,i}^Q \cdot \lambda^{(n-i)}, \quad s_T^Q = \sum_{i=1}^{n} s_{T,i}^Q \cdot \lambda^{(n-i)} \ . \tag{3.2}$$

In the above equations, $\lambda \in [0, 1]$ is a forgetting factor that allows old feedback to be given less weight than more recent feedback, thus, allowing for more dynamic behaviors of agents.

The system also allows to discounting the feedback provided by an agent as a function of the reputation of this agent, following Jøsang (2001)'s work on belief discounting. This way, $\varphi(p|r_T^{X:Y}, s_T^{X:Y})$ is the discounted reputation function of T by X through Y's opinion, such that:

[3] This means that the model is fitted to representations of past experiences that considers different outcomes, although at the end these outcomes must be grouped into binary feedback representing positive and negative evaluations.

$$r_T^{X:Y} = \frac{2r_Y^X \cdot r_T^Y}{(s_Y^X + 2)(r_T^Y + s_T^Y + 2) + 2r_Y^X}, \quad s_T^{X:Y} = \frac{2r_Y^X \cdot s_T^Y}{(s_Y^X + 2)(r_T^Y + s_T^Y + 2) + 2r_Y^X} \,. \tag{3.3}$$

The combined reputation rating of T by his partners is a number that provides an indication of how the agent is expected to behave in the future.

The TRAVOS Model

The TRAVOS (Trust and Reputation model for Agent-based Virtual OrganisationS) model (Teacy et al., 2006; Patel, 2006) is a trust and reputation model for agent-based virtual organizations. It models the trust of a particular agent in a given trustee, by assessing the latter's trustworthiness in a given context based on the truster's previous direct interactions with the trustee and on the opinions of others regarding the trustee. This model intends to minimize bias and errors introduced by others' opinions by judging the manner in which individuals provide opinions and by exploiting social structures. The model assumes that a given trustee $a2$ either fulfills or defaults on its obligations toward truster $a1$ based on his behavior $B_{a1,a2}$, which is given by the intrinsic probability with which the outcome $O_{a1,a2}$ is one, that is, $B_{a1,a2} = E[O_{a1,a2}]$. Hence, the direct trust $\tau^d_{a1,a2}$ at time t is the expected value of $B_{a1,a2}$ given the set of outcomes $O^{1:t}_{a1,a2}$, such that $\tau^d_{a1,a2} = E[B_{a1,a2}|O^{1:t}_{a1,a2}]$. $B_{a1,a2}$ is calculated using the Beta distribution, in a similar way to the one described in the Beta Reputation System, and the incorporation of others' opinions also uses a similar approach based on Beta distributions.

The Model of Reece et al. (2007a)

In (Reece et al., 2007b), the authors propose a formalism based on the Dirichlet distribution that allows to deal with multi-dimensional contracts and avoid the limitations of using the Beta distribution (as, for instance, in the models of Jøsang and Ismail (2002) and Patel (2006)) to only one dimensional contracts. The Dirichlet distribution is used to estimate the probability that each service will be successfully delivered by the trustee and the correlations between these estimates from direct experience of procuring both services.

In Reece et al. (2007a), the authors extend this formalism by allowing the agents to exchange and combine reputation reports over heterogeneous and correlated multi-dimensional contracts. In this extended model, a truster agent is able to fuse his own prior trust estimates about a given trustee

with the reputation report received from a third agent, even when there is not an exact match between the services classified by the truster and the ones reported by the other agent. The combination of these heterogeneous contract observations is made using the Kalman filter, where the missing contract observations are represented by setting the corresponding diagonal elements of the covariance matrix to infinity, or, alternatively, the information form of the Kalman filter, which allows to insert the necessary zeros when these observations are omitted.

The Model of Erriquez et al. (2011)

Erriquez et al. (2011) present an abstract trust framework (ATF) composed of agents and relations of distrust between them, where these agents and relations are represented in a distrust graph. In this framework, any agent may participate in one or more coalitions, or subsets of the ATF, which can be of different types (e.g., distrust free coalition, coalition as a trusted extension of the ATF). In the same way, different types of agents (e.g., trustable agent with respect to a coalition) are formulated. Then, the estimated trustworthiness of a trustee is given by the ratio of the number of maximal trusted extensions of which the agent under evaluation is a member to the overall number of maximal trusted extensions in the system. Also, the coalition expected trustworthiness measures the probability that an agent would be trusted by an arbitrary coalition, picked from the overall set of possible coalitions in the system. Hence, this approach gives special importance to the number of distrust free coalitions for which the trustee under evaluation is a member, independently of the contextual nature of these coalitions. However, intuition tells us that an agent may be engaged in fewer coalitions and still be more trustworthy in one given task and context than other agents that belongs to more, possibly different coalitions.

3.4 Models that Incorporate Trust Dynamics

In Section 2.5, we mentioned that the process of building trust is subject to specific dynamics. For instance, when the trustee under evaluation is a complete stranger, the truster may use third-party information reports, such as opinions, reputation or certificates, to try to estimate the trustworthiness of the trustee (cf. Proposition 28). Some of the computational models that we briefly viewed in the precedent sections present approaches that integrate certified third-party information (cf. the FIRE model), opinions and reputation (e.g., AFRAS, Regret, The Beta Reputation System, TRAVOS).

However, despite his effort, the truster may still not have enough information that allows a trust judgment. Recent computational trust models started addressing the use of social categorization in order to obtain prior information about the trustworthiness of trustees based on their social characteristics (cf. Proposition 21). In this respect, we will overview the models by Burnett (2011), Venanzi et al. (2011), and Teacy et al. (2012) later in this section.

When trust starts growing, some principles are observed, such as the principles of asymmetry and perseverance (cf. properties 31 and 32). The principle of asymmetry of trust is addressed in (Jonker and Treur, 1999; Bosse et al., 2007) and (Melaye and Demazeau, 2005), that we overview below. In turn, to the best of our knowledge, the phenomenon of perseverance is not explicitly addressed by any computational trust approach. The closer we can get regarding this issue is to use the number of experiences to reduce the effect of new experiences when the trustee is already highly reputed, as performed in the SPORAS, AFRAS, and Regret models. In (Hoogendoorn et al., 2009), it is proposed to learn the best values for the different parameters of a given computational trust model that are associated to trust dynamics from the exhibited individual characteristics of trusters. We review this model later in this section.

Finally, when the relationship matures and there are pivotal expectations from each partner of the interaction toward the other partner, these expectations may be violated and the correspondent betrayal may produce severe damage on trust (cf. Proposition 33). To the best of our knowledge, there are no principled computational trust approaches that address the effect of betrayal on trust. A somewhat related, although different concept is *forgiveness*, which embraces the idea that old assessments of a target agent are probably outdated and should not be taken into consideration with the same emphasis as new assessments. This of course does not model the effects of betrayal on trust, because forgiveness implies that the victim of betrayal must act on his diverse negative emotional reactions. Even so, forgiveness can serve as an enabler to restore relationships that would otherwise not be possible (Marsh and Briggs, 2009). Several computational trust models address, through different approaches, the forgiveness property (e.g., Jonker and Treur, 1999; Jøsang and Ismail, 2002; Carbo et al., 2003; Melaye and Demazeau, 2005; Huynh et al., 2006; Marsh and Briggs, 2009). Additionally, Marsh and Briggs (2009) propose regret as a consequence of trust and formalizes the incorporation of this feature in a computational model.

The Model of Melaye and Demazeau (2005)

Melaye and Demazeau (2005) propose a Bayesian trust formalism based on Castelfranchi and Falcone (1998)'s socio-cognitive model of trust. This model uses a Kalman filter to address two dimensions of the trust dynamics: the asymmetric increase/decrease of trust and the inherent speed of switching from trust to distrust and vice versa, named *inertia*; and the erosion of trust that happens due to the absence of new observations. This last dimension meets Hirschman (1985)'s observation that trust grows with use and decays with disuse.

However, the proposed approach seems to scale poorly in the presence of several different beliefs. In the same way, the inertia of trust and distrust is assumed to be fixed a priori by a specialist, requiring one instance of the model per context. Finally, this model seems to be too sensitive to sporadic occurrences of deceptive behavior, as shown in the evaluation of the model (Melaye and Demazeau, 2005).

The Model of Jonker and Treur (1999)

Jonker and Treur (1999) present a framework supporting the analysis and formalization of the dynamics of trust based on experiences. The authors enumerate properties that may be present in trust functions and provide formal specifications of a number of relevant dynamic properties of trust. For example, a computational trust approach may model the predisposition of the truster in the absence of previous trust influencing experiences; a trust evolution function where the number of all positive and negative experiences are counted and compared has the property of indistinguishable past (i.e., if value $+$ represents an event with a satisfactory outcome and $-$ an event with negative outcome, the sequences $+++--$ and $---++$ would yield the some trustworthiness score).

An extended version of this framework is proposed in Bosse et al. (2007). Here, the authors propose an asymmetric trust update function, where negative experiences have stronger impact than positive experiences, as shown in Equation 3.4.

$$\begin{aligned} trust_{t+1}(\delta^+) &= (1-\delta^+)trust_t + \delta^+ \quad \text{if the experience is positive} \\ trust_{t+1}(\delta^-) &= (1-\delta^-)trust_t \quad \text{if the experience is negative} \end{aligned} \tag{3.4}$$

In the equation above, δ^+ and δ^- are the impact factors of positive and negative experiences, respectively, and they are related by an endowment

coefficient e, as shown in Equation 3.5.

$$\delta^+ = e\ \delta^-, 0 < e \leq 1\ . \tag{3.5}$$

The Model of Hoogendoorn et al. (2012)

Hoogendoorn et al. (2009) present an approach to computational trust based on the idea that individuals are different in their characteristics and computational trust models must account for these differences. Ideally, these models must have a number of parameters that are learned for each specific individual based on observed experiences of this individual consulting others (e.g., does the individual ask a human or look in the manual?) and in the outcomes of these consultations. In this paper, the authors consider four parameters: the initial trust value, the decay factor of trust (i.e., how fast the trust decays after a period without experiences), the weight of positive and negative experiences (trust flexibility), and the weight of experiences with competitors upon the trust value (trust autonomy). Several methods have been tested to learn these parameters, including exhaustive search through the space of parameter combinations, Simulated Annealing, bisection, and an extended form of bisection. The authors keep very active in this topic, and an extended version of the initial paper is presented in (Hoogendoorn et al., 2012).

The Model of Burnett (2011)

The model described in (Burnett et al., 2010; Burnett, 2011) addresses the problem of estimating the trustworthiness of trustees in open and dynamic multi-agent systems, where agents frequently join and leave the global population, or the size of the global population prevents the agents to have frequent interactions with known partners. In some way, this is related with the problem of sparse evidence about the trustees under evaluation that we mentioned in our Research Question 5. In order to cope with this problem, the authors propose a stereotyping approach where the agents' observable features provide useful predictors of future behavior for a trust evaluation at the beginning of a trust relationship. In this model, the stereotypes are represented as decisions trees, where each node represents a particular feature. Therefore, traversing the tree using the perceived features of the trustee results in a predicted stereotypical evaluation for the agent. In turn, the resulting stereotype is used when the trustee is a newcomer, and gives an estimated a priori trust value for this trustee.

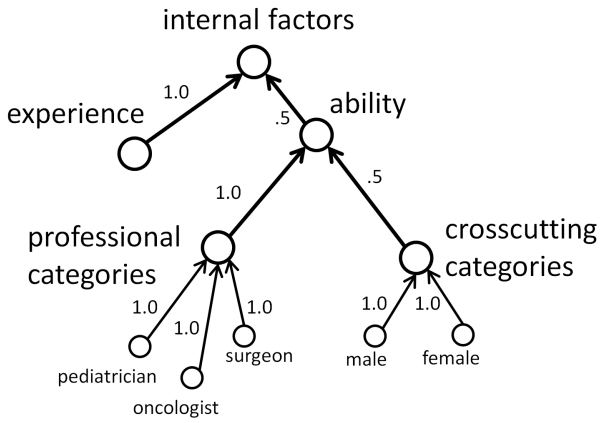

Figure 3.2: Representation of *categorial reasoning* (adapted from Venanzi et al., 2011).

The Model of Venanzi et al. (2011)

In (Venanzi et al., 2011), the authors propose to use *categorial trust* as another trust-based information source of the socio-cognitive model of trust, which will be presented with more detail in Section 3.6. This new source allows to infer hidden information about internal factors of the trustees (the *kripta*) from observable features (the *manifesta*) of these trustees.[4] Figure 3.2 illustrates the portion of the directed graph that represents the process of trust formation in the socio-cognitive model of trust corresponding to categorial trust. This graph is defined for each trustee under evaluation and for each task. Therefore, it requires the configuration of the nodes and weights associated to the professional and crosscutting categories required for a trustee to successfully accomplish the task under evaluation.

The Habit Model

Teacy et al. (2012) present HABIT, a recent probabilistic trust and reputation model. This model defines a parameter vector $\theta_{tr \to te}$ for each truster-trustee pair specifying the distribution that represents how the trustee is likely to behave during an interaction with the truster. For example, for Gaussian distributions, $\theta_{tr \to te} = \langle \mu, \sigma^2 \rangle$, where μ is the distribution's mean and σ^2 is its variance. The agents learn about $\theta_{tr \to te}$ through repeated interaction with *te*, using Bayesian techniques. Then, each truster-trustee pair has a *confidence model* that represents the probability distribution

[4]The authors recover here the idea of kripta and manifesta from Bacharach and Gambetta, as referred in (Castelfranchi and Falcone, 2010).

$p(O_{tr \to te}|\theta_{tr \to te})$ of all observations $O_{tr \to te}$, where $p(O_{tr \to te})$ is a probability measure for possible outcomes of interactions between truster and trustee.

The novelty of the model, though, resides in a second component, the *reputation model*. Here, $\theta_{\cdot \to j}$ is a vector of all parameters used to model trustee j by all known observers, and ϕ is a joint distribution of all parameter vectors for each pair of agents, where each $\theta_{\cdot \to j}$ is independent and identically distributed according to ϕ. Hence, in the HABIT model, a truster performs inference about a specific trustee given observations of any trustee from any source (direct or third party). This means that the model may be used to predict a trustee's behavior based on the behavior of groups of other agents. In a way, this allows to estimate the trust on the trustee even when the latter is a stranger (cf. Proposition 28). However, contrary to other models that use specific physical features of a stranger to infer its trustworthiness (Burnett, 2011; Venanzi et al., 2011), the HABIT model predicts this trustworthiness based on the most common observed behavior of other trustees in the particular situation under assessment, by learning ϕ, which does not take into account the individual differences of each trustee. Alternatively, the authors suggest to partition agents into groups with similar behavior (e.g., by using cluster algorithms to attributes relevant to the specific application domain, as proposed by Burnett, 2011) and to have a separate reputation model for each group.

Finally, in order to produce situation-aware trust estimations (cf. Proposition 1), a truster must maintain a different confidence model of each trustee per context of interest. Every observation out-of-context is treated as if it was reported by a different observer and then the reputation model is used to learn the correlations between observations from different contexts. In our opinion, this approach has a major limitation: it reduces the number of direct observations of the trustee used in the confidence model. This evidence is usually sparse and should not be transferred to the reputation algorithm, because by doing that important information about the benevolence of the trustee is lost and in consequence the estimation of the trustee's trustworthiness results less accurate (Proposition 10).

3.5 Situation-aware Trust Models

As trust is situational (Proposition 1), the need for some kind of computational situation-aware trust is evident. In fact, it is realistic to assume that not all past evidence is equally relevant to assess the success of future interactions, as it is common sense that a given entity might behave differently in different social contexts. As an example, a report saying that John

is a good cook is almost useless when estimating how trustworthy he is in driving his friends back home in safety. This same reasoning applies to the inference of trustworthiness in social graph-based structures, much used in reputation based and recommendation systems. In fact, although existing models of reputation are generally based on the transitivity of trust, it is wise to note that trust is not always transitive (Christianson and Harbison, 1997) and that extra care is needed to incorporate a situational dimension into graph-based reputation models (Tavakolifard et al., 2009).

The consideration of context can also help to reduce the complexity inherent to managing trust relationships (Neisse et al., 2007), as well as to bootstrap unanticipated situations, where missing information on a trustee can be inferred from similar situations; for instance, if we know that John is a proficient piano player, we can use this information to estimate the trustworthiness of John as a piano teacher (even though John may be a lousy pedagogue). Other uses of situational trust include domains where the agents perform diverse tasks in highly dynamic environments, wireless sensor networks where possible interactions depend heavily on the context domains, network intrusion detection, and ubiquitous computing (Rehák and Pěchouček, 2007; Rehák et al., 2008; Tavakolifard et al., 2008). Despite the importance of context in trust, only a few computational trust models allow to make context-based trustworthiness estimations, most of them are based on ontologies.

In Regret (Sabater and Sierra, 2001, cf. Section 3.3), a given trustee is assessed in a given scenario, and this scenario is constituted by different aspects organized in an ontological structure. The reputation score of this trustee is calculated as the weighted means of the trustee's reputation in each one of these aspects. However, these aspects may be themselves contextual (e.g., a trustee that usually delivers a given product on time may tend to delay the delivery of other products or the fulfillment of other types of agreements), which may reduce the effectiveness of this approach.

In (Toivonen and Denker, 2004), the authors propose an approach in the domain of message-based communications that creates trust policies using rules based on the explicit context of messages in order to determine the trustworthiness of these messages. In order to capture the message content and the context-dependent trust relations, the authors extend a trust ontology proposed by (Golbeck et al., 2003). As an example, a given trust policy may state that the trustworthiness of a message reporting an event at a given location is higher if the reporting entity was at this location at the time of occurrence of the event.

A different approach, inspired in the research area of collaborative fil-

tering, is given in (Tavakolifard et al., 2009; Tavakolifard, 2009). It uses taxonomy-based similarity measures to derive the similarity between users and classes of items they tend to share (also known as co-rating behavior). We present this approach with a little more detail later in this section. Nakatsuji et al. (2010) proposes a related approach in the research area of cross-domain recommendations where users that share similar items or social connections provide recommendations chains on items on other domains, using Web taxonomies made available by service providers. In order to allow measuring the similarity of users that do not have rated the same items, the model first computes class classifications from the individual classifications of items of the class, and then computes the similarity between users taking into account the resulting similarities of users in each class. This model implies, however, that certain subtleties that may exist between items within classes cannot be taken into account because of the generalization process it assumes.

Hermoso et al. (2009) propose a model of trust that uses as information sources the direct experience with the trustee and also information provided by organizational roles (Proposition 21), where the role taxonomy is dynamically updated from trust information maintained by the agents using clustering techniques. When evaluating the trustworthiness of a target agent, the evaluator uses the direct experiences he has with the trustee and weights them according to the similarity between the role assumed by the agent in the specific experience and the role that is assumed in the current situation.

Other approaches that allow for some sort of context representation using ontologies are presented in (Jung, 2008; Fabregues and Madrenas-Ciurana, 2009). Although the use of taxonomies and ontologies is increasing in the Web, both in social networks and e-business activities, the computational trust approaches that use these approaches to model context are constrained by the necessity to predefine adequate similarity measures for all possible situations in assessment before such situations are even presented to the evaluator. This is a domain specific, hard tuning process that may be a challenge in dynamic environments with complex representations of contexts. Also, there are subtleties in situations that may not be well captured by hierarchical-based similarity measures. As an hypothetical example, these models may uncover that the situation delivery of one container of *cotton* from Asia to Europe is quite similar to the situation delivery of one container of *chiffon* from Asia to Europe, but they could fail to discover that the trustee under evaluation tends to fail these deliveries in the presence of short delivery times.

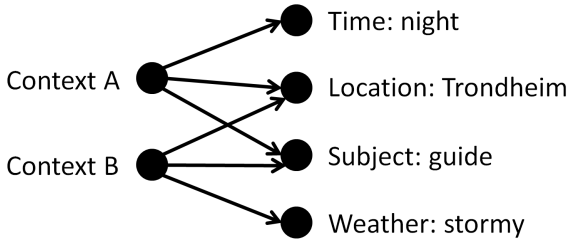

Figure 3.3: Relation between contexts and aspects in Tavakolifard et al. (2008)' model.

The Context Management Framework Model

The Context Management Framework (CMF) model (Tavakolifard et al., 2008, 2009; Tavakolifard, 2009) uses case-base reasoning techniques to estimate the trustworthiness of agents in unanticipated situations, by retrieving the most similar cases to the unanticipated situation from a case base. In order to represent the similarity, the model uses a context-specific trust ontology and measures of relational similarity, which are based on the SimRank algorithm (Jeh and Widom, 2002). Figure 3.3 illustrates a practical use of this algorithm. Context A refers to Alice trusting Bob to guide her in Trondheim at night, and Context B concerns Alice trusting Bob to guide her in Trondheim when it is stormy. Contexts A and B share two aspects: the location (Trondheim) and the subject (guide). The SimRank algorithm is based on the assumption that, in a general way, two contexts are similar if they are related to two aspects that are themselves similar. Hence, it is possible to derive the similarity between Context A and Context B through the aspects they share (i.e., location and subject).

The major drawback of this approach resides in the weak assumption made by the SimRank algorithm about the similarity between different objects. As the authors recognize (Tavakolifard, 2009), the similarity of two context models is itself context dependent, preventing the model to adequately scale to more complex representations of contexts. For instance, context A and B could share all aspect but the subject, and this would be sufficient to alter significantly their similarity relation.

The Context Space and Reference Contexts Model

The Context Space and Reference Contexts model (Rehak et al., 2006; Rehák and Pěchouček, 2007; Rehák et al., 2008) assumes that the context space is Q-dimensional and has one dimension q per relevant feature of the

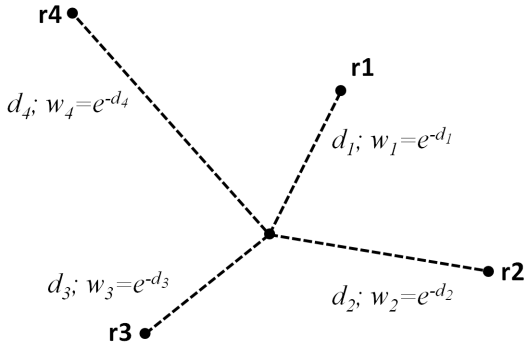

Figure 3.4: The context space (adapted from Rehak et al., 2006).

environment. Also, for each considered dimension, the model defines an appropriate distance metric $d(c1, c2)$ describing the similarity between any two contexts $c1$ and $c2$. Figure 3.4 illustrates this contextual representation.

In this model, each truster must maintain a context space for every trustee (or group of trustees with similar characteristics) under assessment. In each one of these context spaces, the truster places $n < Q$ *reference contexts* (either regularly or adaptively). In the presence of new evidence about the trustee under assessment, the trustworthiness at each one of the reference contexts of this agent's context space is updated with the outcome of the evidence, weighted by the similarity between the reference context and the context of the new evidence. Then, when the truster needs to assess the trustee's trustworthiness in a specific situation, the model uses the most similar reference contexts and the trust score is computed by using a weighted means of the trustworthiness values at these reference contexts weighted by the similarity between the new situation and these reference contexts. Equation 3.6 illustrates the computation of the trustfulness of the trustee at context reference r_i when a new individual item of evidence $(p+1)$ of the trustee's trustfulness in situation co is available. In the equation, W^p is the aggregate weight of previous (p) observations and ω^{p+1} is a function of the distance between reference context r_i and the situation co concerning the new observation.

$$tw_{r_i}^{p+1} = \frac{W^p \times tw_{r_i}^p + \omega^{p+1} \times tw_{co}}{W^p + \omega^{p+1}}, \quad \omega^{p+1} = e^{-d(r_i,co)} \ . \tag{3.6}$$

This is an interesting model with some attached limitations. First, as we have mentioned, this model depends on predefined measures of similarity between contextual attributes, which can make the configuration process cumbersome in rich contextual scenarios. Besides, the model does not ad-

dress the fact that the distance metrics may be themselves contextual. Also, the consideration of multiple dimensions can lead to an exponential number of reference contexts that each truster needs to maintain for every trustee, jeopardizing the scalability of the model in complex contextual scenarios. At the same time, the performance of the model may be reduced if the evidence on the trustee under evaluation is sparse and distributed over many reference contexts. In order to minimize these negative effects, the authors propose to use clustering algorithms to allow for a dynamic placement of the reference contexts, avoiding the creation of reference contexts for unusual situations.

Nguyên and Camp (2008) present a related approach to represent context that is used together with a Bayesian trustworthiness aggregation engine. In this approach, each dimension of the context space is now a function of many context attributes, instead of one single attribute. However, all limitations pinpointed above for the previous model are still present in this approach.

The Socio-Cognitive Model of Trust

The socio-cognitive model of trust of Castelfranchi and Falcone (2010) considers the situation-awareness of trust assessments in two distinct perspectives: the *evaluation context*, which shapes the trust evaluation and decision of the truster, affecting his mood, social disposition, risk perception, beliefs activated, sources and information used, etc; and the *execution context*, which affects the objective trustworthiness of the trustee. In this last perspective, the truster must have a perception of how beneficial or harmful the environment where the task is to be executed can be to the trustee, and what is the influence of the supporting infrastructure, institutional context, and generalized social values. The final trust decision results from the combination of the trust on the trustee (internal attribution about the perceived competence and disposition of the trustee) and environmental trust (external attribution, or evaluative beliefs about the contextual environment, including opportunities of the trustee to realize the task, corresponding to the execution context mentioned before), as illustrated in Figure 3.5.[5]

We notice a different approach to situation-awareness in this model, when compared to the ones we briefly viewed before. In fact, the other models tend to look to past evidence and reason about the most probable behavior of the trustee in current situation based on past interactions in similar situations. In the socio-cognitive model of trust, the truster has a

[5]This figure is part of a broader model of social trust based on goals and beliefs that we present in Section 3.6.

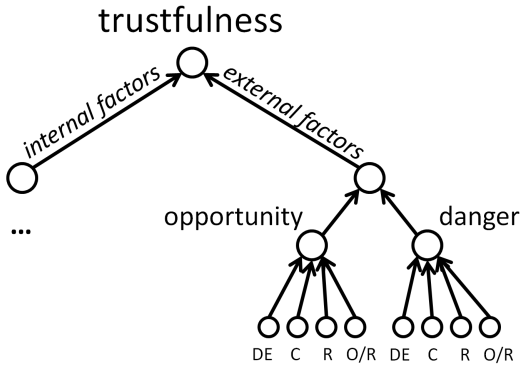

Figure 3.5: The role of context in the Socio-Cognitive Model of Trust (adapted from Castelfranchi and Falcone, 2010).

set of beliefs about the opportunities and dangers inherent to the trustee performing that task at this given moment and time, where the focus is more on the context itself and less on the trustee under evaluation.

Although theoretically interesting, the author's view about situation-aware trust is only partially addressed in the current implementation of their model, and an explicit representation of context seems to be lacking, as well as a formalization of the notion of evaluation context. In fact, in Venanzi et al. (2011), the environment influences are modeled as a parameter assuming values in $[-5, +5]$. In the same way, the current implementation of the model suggests that the evaluation of a trustee in two distinct situations requires the maintenance of two different hierarchies of beliefs, where both the bottom beliefs nodes and the causal power of basic beliefs would be different (Castelfranchi et al., 2003; Venanzi et al., 2011), which requires additional configuration effort and computational complexity.

3.6 Social-based Models of Trust

Some of the earliest computational models of trust and reputation that we mentioned in Section 3.3 present some characteristics of social trust. For instance, the AFRAS model (Carbo et al., 2003) presents an architecture of agents with three layered models: world, social and mental. The top layer is the mental layer, which manages different attributes of the agent, such as shyness, egoism, susceptibility and remembrance, whose values affect the inferior layers, responsible for deciding with whom to interact. Hence, this model has concerns about the way that the dispositional traits of agents affect their social interactions. However, the modeling of the agents' attri-

butions is rigid and oversimplified. For example, agents that act socially are modeled to interact with newcomers; agents that are egoist do not answer to questions posed by others; and agents that are shy do not pose questions to others. In the Regret model (Sabater and Sierra, 2001), there is a notion of social reputation that is based on the authors' claim that "belonging to a certain group implies, a priori, that its members share a common way of thinking"; as we have seen in Section 3.3, group reputation may provide initial expectations about the behavior of an agent when direct information from personal interactions with this agent is lacking.

However, despite these episodic notes on social trust, a deeper understanding of the social account of (computational) trust is needed. We review the most significant of these approaches until the end of this section.

The TRAVOS-R Model

Patel (2006) present TRAVOS-R, an extension to the TRAVOS model (cf. Section 3.3) that aims at incorporating the knowledge of social relationships existing between agents and to subsequently use it in the calculation of trust. For this, the TRAVOS-R model assumes that relationships belong to a particular type (e.g., cooperative, competitive, and dependence), and that two agents are related to each other by a permanent relationship of a certain type, even though transient relationships between the two agents may exist in different periods of time. Moreover, the model assumes that agents maintain a conditional probability table specifying the probability of a certain action being observed, given the fact that a certain type of relationship is present. Hence, by observing the actions for all types of observed transient relationships and using the Bayes rule, the agents are able to calculate the posterior probability for each type of relationship and calculate the type for the permanent relationship. Finally, a truster that uses an opinion about the trustee under evaluation may discount or compensate the value of this opinion taking into consideration what he knows about the relationship between the opinion maker and the trustee. Thus, the TRAVOS-R model presents an approach to evaluate the relationships existing between the trustees and their evaluators, which in part corresponds to the Research Question 3 that we formulated in the introductory chapter of this thesis.

This approach presents, however, several limitations. First, it assumes the existence of predefined relationship types and actions. Second, it assumes that two related agents have a permanent relationship of a certain type that is static, even though different transient relationships may develop. Moreover, it also assumes that the actions produced by agents in

their interactions are not a function of the previous actions of these agents. Third, it is supposed that trusters acknowledge several interactions between the trustee and the opinion makers, in order to maintain the posterior probability for each type of relationship based on the observed actions, and that they are aware of all actions performed in these interactions. In this case, if the truster has access to all this detailed data, he probably has information enough to infer the trustworthiness of the trustee instead of relying on the opinions provided by the third-party agents. Fourth, by stating that competitive agents tend to falsify opinions and cooperative agents tend to provide exaggerated opinions, the model does not account for the integrity (or even benevolence) dimension of agents.

The Model of Adali et al. (2011)

Adali et al. (2011) present a model for social trust that accounts for a vast part of the thematic of trust and trustworthiness that we addressed in Chapter 2. For instance, this model considers that the estimation of the trustee's trustworthiness is accomplished through perceptual clues, social clues, and information derived from past experiences; and that the trust decision takes into account the trustee's trustworthiness, the *social trust* (including reputation and recommendations), the trust propensity of the truster, and the context of the trust situation.

This model is based on the Kelton et al. (2008)'s integrated model of trust that we described in Section 2.3.6, and integrates a broader approach to the modeling and computation of trust in composite networks. In the authors' view, trust decisions must take in attention the dependence existing between the social (trust in persons), informational (trust in information sources) and systemic (trust in systems) networks. These different networks are represented as factors of trust, and any truster must decide at each situation what factors to select and how to combine the trust assessments generated by the selected factors. The impact of each factor on the trustworthiness of a trustee is represented probabilistically in terms of the mean probability and a measure of confidence in that probability.

Although conceptually grounded, the social account of this model does not appear to be fully implemented. In fact, the details about the representation of the truster disposition to trust and the trustee's social and physical clues, as well as the way they combine, are not provided. Moreover, the authors state that more sophisticated realizations would be needed to treat the cognitive and social concepts involved, as well as the utilities and economic preferences. Finally, the authors delegate to future work the deeper study

of how social norms contribute to trust and how the notion of social trust can be inferred from the network structure.

The Socio-Cognitive Model of Trust

The socio-cognitive model of trust was first proposed by Castelfranchi and Falcone (Castelfranchi and Falcone, 1998) and later developed and implemented within their research group (e.g. Castelfranchi et al., 2003; Castelfranchi and Falcone, 2010; Venanzi et al., 2011).

The authors consider that trust is (at least) a five-part relation, where a truster X trusts a trustee Y in a context C for performing a task α that will result in an outcome that includes or corresponds to the goal g of X. They claim that only a cognitive agent endowed with goals and beliefs can trust another agent, and that to trust implies to trust relative to a goal (a need, a desire, an achievement). Moreover, trust is a mental attitude based on beliefs about the trustee's internal features, such as the beliefs about the ability, disposition and unharmfulness of the trustee in performing the needed action, and on beliefs about external attributions, such as opportunities, resources, interferences and adversities.

Every belief is formed from belief sources of four different types: direct experience (DE), categorization (C), reasoning (O/R), and reputation (R). For example, a truster may have a belief about the ability of a given trustee based on reputation. In turn, this belief results from several individual opinions, or single beliefs, about the ability of the trustee (e.g., "John says that the doctor is quite good at his work"). Figure 3.6 illustrates the process of formation of belief sources from single beliefs, modeled as a direct graph. The impact of the single belief on the belief source is given by the value of the content of the belief (ascribed to the belief node), which is subjectively modulated by epistemic evaluations about the source of the belief, namely, the certainty about the source (how credible is John), the source's subjective certainty about the belief (e.g., how John beliefs in what he have communicated), and the truster's trustfulness about the source (how the truster beliefs in the fact that John have said what he communicated). This modulation is used to derive the weight of the edge that connects the primitive belief to the belief source.

The authors provide a possible implementation of their model using Fuzzy Cognitive Maps (FCM).[6] The FCM allows for a direct implemen-

[6] A FCM is a graph where processes are modeled by nodes and the causal relations between processes are modeled by weighted edges. The value at a node/concept is given by a function of the sum of the values of the incoming nodes weighted by the values of the corresponding edges.

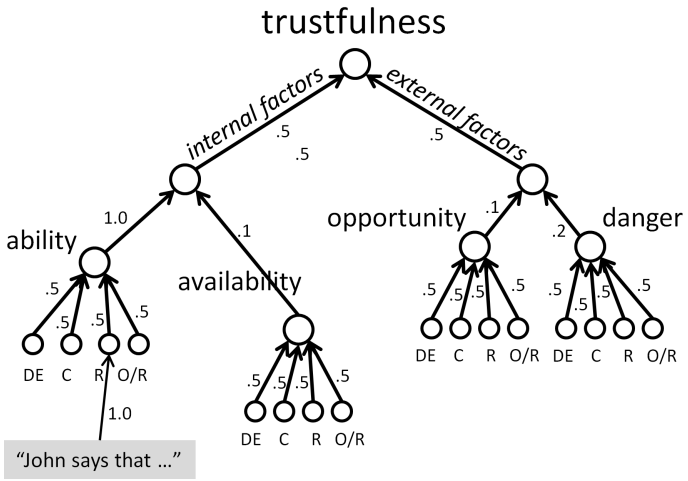

Figure 3.6: Formation of belief sources from single beliefs (adapted from Castelfranchi and Falcone, 2010).

tation of the four layer directed graph illustrated in Figure 3.6. It is still possible to nest other FCM's to implement the process of formation of the beliefs at the leaf level from individual other beliefs. Also, theoretically, the weights of the edges of the FCM may be configured in order to reflect the trust disposition of the truster.

We devoted a great number of lines to describe this model because it is the result of more than one decade of research on social-based trust. This model is well grounded and rich in terms of cognitive construction, presenting insightful epistemological considerations about the sources of the beliefs. However, the richness of the model makes it hard to implement in practice. In fact, the current FCM-based implementation has some limitations and assumes several simplifications to the conceptual model. On the one hand, the model requires a great amount of configuration, as every FCM layout must be designed by domain experts, and every truster must maintain different FCM per trustee *and* task. Also, capturing the mutual influences between beliefs may be a very complex task, even for experts. On the other hand, the model requires extensive information about beliefs and meta-beliefs that may be hard to get in dynamic agent-based systems. As an example, the information about the current dangers and opportunities associated to the trustee realizing a specific task in a given situation may not be available. And the same applies to the empirical data needed to make the epistemic considerations about the belief sources providing the existent information. Finally, the impact of the beliefs on upper beliefs in the current implemen-

tation is fixed, instead of reflecting the desired epistemic evaluations about the source of the belief. Hence, the mentioned constraints and simplifications make it less clear whether the current practical implementation of the model still keeps the virtues of the theoretical model.

The Model of Herzig et al. (2010)

Herzig et al. (2010) formalize the socio-cognitive model of trust described in the previous subsection. In this formalization, it is given special emphasis to the aspect of motivation: trust involves the truster to have a goal, and it is this that allows to distinguish trust from mere thinking and foreseeing. Therefore, this model reduces trust to the more primitive concepts of belief, goal, capability, and opportunity: a truster i trusts j to do α in order to achieve φ, if and only if i has the goal φ, i believes j is capable to do α, i believes that j, by doing α, will ensure φ, and i believes that j intends to do α.

This model also adds to the socio-cognitive model of trust a distinction between *occurrent trust*, i.e., trust in the occurrence of the action α "here and now", and *dispositional trust*, i.e., trust in a general disposition of the trustee to perform a given type of action. This dispositional trust does not have a direct match in the propositions that we have derived in Chapter 2. However, we could think that this disposition is the result of the joint effect of the ability (to perform the type of action), integrity (to try the best to be successful in any action that results from an agreement) and benevolence (the willingness to do good to the particular truster) of the trustee. These concepts of occurrent and dispositional trust are formalized in a multimodal logic L developed by the authors.

3.7 Computational Reputation Models

Although the field of computational reputation has its own set of research questions and challenges, different academics have proposed models of computational trust and reputation that integrate both social concepts, assuming the perspective of reputation as an antecedent of trust (e.g., Abdul-Rahman and Hailes, 2000; Banerjee et al., 2000; Sabater and Sierra, 2001; Jøsang and Ismail, 2002; Yu and Singh, 2002; Carbo et al., 2003; Jurca and Faltings, 2003; Ramchurn et al., 2004; Maximilien and Singh, 2005; Sabater-Mir et al., 2006; Patel, 2006; Huynh et al., 2006; Jøsang et al., 2007; Salvatore et al., 2007; Mundinger and Boudec, 2008; Teacy et al., 2012). Most of these models estimates the reputation score of a given trustee from the ag-

gregation of reputation ratings from various inputs, although a propagation mechanism designed to collect the different reputation values may also be considered a mandatory component of these models.

The interplay between trust and reputation raises different challenges, which are inherent to the unreliable nature of reputation. We present next some of the most relevant challenges, which are derived from the computational reputation literature mentioned above. First, computation trust models that use reputation need to estimate the credibility of both the transmitted information and the agents reporting the information, in order to weight the received information accordingly. This estimation must take into account the possible change in the quality of opinions of information providers. More advanced models would be able to determine the kind of agent to obtain reputation information from. Second, computational reputation models must know how to access the opinion makers. Third, these models must provide adequate incentives for referrals to provide reputational information. Fourth, these systems must be able to tackle the problem of the heterogeneity of the different images that constitute the reputation score being transmitted, both at the syntactic and the semantic level.

3.8 Concluding Remarks

In this chapter, we gave a view of different aspects of computational trust and described with some detail the computational trust approaches that we consider most relevant in the scope of this thesis' work. In this view, we saw that all these approaches implement an aggregation algorithm allowing to estimate a trustworthiness score based on the evidence available on the trustee. A great number of these approaches are dedicated to the management of reputation and estimation of reputation scores (cf. propositions 24, 25, 26), and hence they present simplified models for the computational trust function. In order to better analyze the state-of-the-art of computational trust concerning the social account of trust, we compared these approaches against the theoretical propositions that we derived in the previous chapter. For the sake of systematization, we condensed the results of this analysis in Table 3.1.

The main conclusions that we obtained from this study is that the majority of the analyzed computational trust approaches fails to model social trust. In fact, most of these approaches model trust as a cognitive process that leads to the estimation of a final degree of trust (Proposition 5) that reflects a probability of success in a future interaction with the trustee under evaluation. The behavioral account of trust is not explicitly assumed

in these approaches (Proposition 4). In the estimation of this probability of success, the emotional content of trust is not addressed at all, and then propositions 2, 3, 29, 30 are not addressed. In the same way, most of these approaches do not consider other antecedents to trust than trustworthiness (and sometimes reputation, Proposition 25), failing to address propositions 6, 20, 22, 23. Moreover, only two of these approaches consider that trustworthiness is a multi-dimensional construct, and only one addresses in practice the role of ability, integrity and benevolence in the formation of the trustee's trustworthiness. Thus, all the other approaches fail to consider propositions 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19.

Concerning other properties of the dynamics of trust, the majority of the trustworthiness estimators studied does not account for the asymmetry of trust (Proposition 31) and none addresses neither the perseverance of trust nor the effects of betrayal on trust (propositions 32 and 33, respectively). In the same way, only some of the most recent approaches propose to use social categorization to try to estimate the trust on strangers (propositions 21 and 28). Finally, there are a number of computational trust models that address the situation-awareness in trust assessments (Proposition 1), most of them use ontology-based distance functions to measure the similarity between two different situations.

In the next chapter, we formulate the main hypothesis of our work, derived from the research questions introduced in Chapter 1 and the knowledge we acquired in this and the precedent chapter. Then, we present the SOLUM model, our approach to computational (social) trust. In the description of SOLUM, we make explicit references to the theoretical propositions that are addressed by our model.

Table 3.1: Comparison of different computational trust and reputation models based on the propositions about social trust derived in Chapter 2.

#	Proposition	Marsh (1994)	SPORAS	AFRAS	Regret	FIRE	Beta Rep. Sys.	TRAVOS	Reece (2007)	Erriquez (2011)	Melaye (2005)	Jonker (1999)	Hoogend. (2012)	Burnett (2011)	HABIT	CMF	CSRC	TRAVOS-R	Adali (2011)	Socio-Cogn.
1	Trust as Quaternary Relation	x			x									x	x	x	x	x	x	x
2	Cognitive and Emotional Content																			x
3	Effect of Positive Emotions																			x
4	Behavioral Context																			x
5	Degree of Trust	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
6	Antecedents to Trust																		x	x
7	Trustw. as Multi-Dimensional																		x	x
8	Ability																		x	x
9	Perception of Ability																			x
10	Importance of Benevolence																			
11	Dispositional Benevolence																			
12	Benevolence with Alike																			
13	Relational Benevolence																			
14	Satisfaction with Relationship																			
15	Mutualistic Benevolence																			
16	Integrity																		x	x
17	Perception of Integrity																			
18	Percep. Indiv. Trustw. Dimensions																			
19	Rel. Imp. Trustw. Dimensions																			

continued ...

...continued

#	Proposition	Marsh (1994)	SPORAS	AFRAS	Regret	FIRE	Beta Rep. Sys.	TRAVOS	Reece (2007)	Erriquez (2011)	Melaye (2005)	Jonker (1999)	Hoogend. (2012)	Burnett (2011)	HABIT	CMF	CSRC	TRAVOS-R	Adali (2011)	Socio-Cogn.
20	Propensity to Trust																		x	x
21	Effect of Social Categorization													x						x
22	Effect Kinship/Cultural relatedn.				x															
23	Effect of Emotional State																			
24	Reputation																			
25	Reputation as Anteced. Trust	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
26	Cred./Relevance Trust Sources																			
27	Ignorance and Uncertainty						x	x												
28	Trusting Strangers																			
29	Trustin In-Between Relationships																			
30	Trust in Ongoing Relationships																			
31	Asymmetry										x	x	x							
32	Perseverance																			
33	Trust after Betrayal																			

Chapter 4

SOLUM – Situation-aware and Social Computational Trust Model

In this chapter, we present SOLUM – Situation-aware and sOcial computationaL trUst Model, our agent-based approach to computational social trust, which takes into consideration the properties we identified as vital in Chapter 2 and underpins the contributions of the thesis.

The SOLUM model is composed of two distinct parts. The first part consists of a general computational trust framework constituted by different evaluation functions, based (essentially) on the proposition that trust is a multidimensional construct (Proposition 6) and that the trustees' trustworthiness may be roughly built upon the ability, integrity, and benevolence dimensions of trustworthiness (Proposition 7). The second part of the SOLUM framework consists of a set of computational components that we have developed as possible instantiations of the frameworks's evaluation functions. These components are *Sinalpha*, *Contextual Fitness*, *Social Tuner*, and *Integrity Tuner*. We also developed a method for combining the estimated values of ability, integrity and benevolence into a trustworthiness score, taking into consideration the current stage of the relationship between the truster and the trustee under evaluation (propositions 18 and 19).

4.1 Introduction

In Chapter 1, we formulated the research questions that guided the early stages of this PhD work. Then, in chapters 2 and 3 we reviewed the background about social trust and computational trust that allowed us to revise

these early research questions. From this study, we focused our attention on some fundamental questions. First, most computational trust approaches estimate the trustees' trustworthiness using individual items of evidence about these trustees' behavior in past interactions, either with the truster or with third party agents. However, none of these approaches is able to estimate from the set of past evidence what the benevolence of the trustee toward the truster evaluating him is. As we have been mentioning throughout this thesis, the particular outcome of an exchange may depend not only on the ability and integrity of the trustee, but also on the benevolence relationship that exists between the latter and the truster. Also, we believe that understanding the benevolence of the truster toward the trustee is fundamental to an accurate estimation of this trustee's trustworthiness.

Therefore, we are in conditions of presenting the first hypothesis of this work, as follows.

Hypothesis 1 *The extraction of benevolence-based information from the set of evidence on the trustee under evaluation and its use in adequate stages of the relationship between truster and trustee improves the reliability of the estimation of this trustee's trustworthiness.*

Another question is related with the contextual nature of the trust construct in general (Proposition 1) and of the ability and integrity dimensions of trustworthiness, as referred to in proposition 8 and 16. We have seen that only a few computational trust approaches address context, and these tend to rely on ontologies or other similarity-based predefined distance functions, which we consider may not bring the necessary flexibility in very dynamic agent-based systems where the evidence about a given trustee may be scarce. We believe that a different approach more focused on patterns of past behavior of trustees may be more adequate to address situation-awareness in trustworthiness estimation, in these conditions. This impelled us to formulate the second hypothesis of our work, as follows:

Hypothesis 2 *The use of proper computational techniques that enable to extract contextual information from the set of evidence on the trustee under evaluation improves the reliability of the estimation of this trustee's trustworthiness. The consequent reliability of the trust decision is improved even when the available evidence is scarce.*

From our analysis of the existent computational trust approaches that we summarized in Chapter 3, we verified that most trustworthiness estimators are based on some measure of central tendency and thus are not truly

sensitive to the consistency of the trustees' behavior. However, there is a relative broad consensus in the integrity literature on defining a person of integrity as a person that remains true to his relevant acts and then is consistent in his outcomes. Hence, we wonder if measuring the consistency of the outcomes of the trustee under evaluation from all evidence available on him may improve the reliability of the estimation of this trustee' trustworthiness. We formulate, then, the following hypothesis:

Hypothesis 3 *The extraction of integrity-based information from the set of evidence on the trustee under evaluation improves the reliability of the estimation of this trustee's trustworthiness.*

The final question that draw our attention was related to the work that we are developing in the LIACC Laboratory on the ANTE platform. We verified that while the joint use of trust and normative control has received large attention from the scientific community, mostly in the context of intra- and inter-firm relations, very little is being done in the proposal of computational solutions binding trust and norms. Concerning this issue, we aimed at verifying if the proficuous bind between both governance mechanisms as described in the literature could be easily translated into a computational solution. Based on this, we are in conditions of formulating our final hypothesis, as follows:

Hypothesis 4 *The use of trust in contracting agent systems decreases the weight of sanctions without jeopardizing the efficiency of normative control in promoting the compliance of agents.*

In order to test the first three hypotheses, we built the SOLUM framework, a general framework for building trust models that takes into consideration the main trust-based propositions derived previously. This framework is composed of separated functions allowing for the evaluation of the ability, integrity and benevolence of the trustees under evaluation; the evaluation of the general trustee's trustworthiness given these individual dimensions of trustworthiness; and the evaluation of the trust the truster has on the trustee given his trustworthiness and other factors. The SOLUM framework is presented in Section 4.3.

In the same way, we propose a particular instantiation of the SOLUM framework by developing computational components that implement each one of the functions that estimate the trustworthiness dimensions. These components are *Sinalpha* (Section 4.4), *Social Tuner* (Section 4.7), and *Integrity Tuner* (Section 4.6). We also present *Contextual Fitness*, a component that extracts contextual information from the set of evidence on the

trustee under evaluation, allowing for situation-aware assessments (Section 4.5), and a general method for combining the estimated values of the trustworthiness dimensions taking into consideration the context and the stage of the relationship between the truster and the trustee under evaluation (Section 4.8).

The last hypothesis formulated earlier in this section is tested using the ANTE platform, referred to in Chapter 1. We settle an environment where trusters had the choice to select partners based on either trust, sanctions, or both. The ANTE platform have also allowed us to perform additional studies on the use of trust in negotiation processes. All work related to trust as a service of the ANTE platform, including the any experimental analysis we conducted, is presented in Chapter 6.

Next, before we proceed into the description of the SOLUM framework, we present a set of basic notation that we shall use throughout the remainder of this thesis, including our own representation of context.

4.2 Basic Notation

We assume the existence of a society of agents represented by the limited set $\mathcal{A} = \{a_1, a_2, ..., a_n\}$. In this society, agents make trust decisions about other agents concerning the realization of a given task $t_i \in \mathcal{T}$ in a given situation $s_i \in \mathcal{S}$, where $\mathcal{T} = \{t_1, t_2, ..., t_m\}$ is the set of all possible m tasks in the society and $\mathcal{S} = \{s_1, s_2, ..., s_k\}$ is the set of all possible k situations in the society.

We are interested in the interactions of any pair of agents $(a_1, a_2) \subseteq \mathcal{A}$ governed by agreements specifying the obligations of each agent toward its interaction partner. These agreements may be informal or formal, and in this last case we name them contracts. Furthermore, each one of the interacting agents may totally fulfill the ascribed obligations, partially fulfill them, or totally violate them. Therefore, each one of these partners trusts the other to some degree to perform his obligations, and when the interaction is over each one of them is able to indicate the perceived outcome $o_i \in \mathcal{O} = \{o_1, o_2, ..., o_p\}$ of the interaction from his perspective. It is evident from now that as both agents perform at the same time the role of truster and trustee, two outcomes o_{a_1,a_2} and o_{a_2,a_1} will be generated. We provide more information about the set of all possible outcome values $\mathcal{O}$ in subsection 4.2.2. For convenience, we further represent an agent playing the role of a truster by $x \in \mathcal{A}$ and an agent playing the role of a trustee by $y \in \mathcal{A}$. Then, in a given point of time and situation, $\mathcal{X} \subset \mathcal{A}$ is the set of all existing agents playing the role

of trusters and $\mathcal{Y} \subset \mathcal{A}$ is the set of all existing agents playing the role of trustees, and $\mathcal{A} = \mathcal{X} \cup \mathcal{Y}$.

Furthermore, we consider in our model that the evidence available on any given trustee is derived from the agreements that this trustee established with others (including the truster that is currently evaluating him) in the past. Hence, each individual item of evidence $e_i \in \mathcal{E}$ has a correspondent outcome $o^{e_i} \in \mathcal{O}$ that presupposes an evaluation of the trustee by his interacting partner in the agreement, where $\mathcal{E} = \{e_1, e_2, ...\}$ represents the set of all individual items of evidence produced in the society to date. The set of all evidence on trustee y is represented by $E_{*,y}$, with $E_{*,y} \subseteq \mathcal{E}$. This set includes the set of all evidence on y concerning the direct interactions of this agent with truster x, $E_{x,y}$ (i.e., $E_{x,y} \subseteq E_{*,y}$). At this time, we say that agents x and y had interacted $N_{x,y}$ times.

Our representation of evidence just described implies two different things. First, any individual item of evidence e_i that the truster may have about the trustee – and that will be used to trustworthiness inference – may be obtained through direct interaction with the trustee (when $e_i \in E_{x,y}$) or indirectly, through information obtained from other trusters or third-party certificated entities (when $e_i \in E_{*,y}$ and $e_i \notin E_{x,y}$). However, it always relates to an agreement and makes reference to the outcome of this agreement. Therefore, it does not include reputational information of any form, neither general assessments about the qualities of the trustee (e.g., 'the trustee is very able in performing this task', or 'the trustee is a person of integrity'). Although this last type of information is valuable, it is hard to obtain and to integrate, as we have already discussed in Section 3.6 when reviewing the socio-cognitive model of trust. We consider to integrate this other type of information in future versions of our model, using, for instance, a graph-based solution as considered in (Castelfranchi and Falcone, 2010).

Second, we assume that the set of evidence that the truster has on the trustee under evaluation is correct, even if acquired through third-party entities. As we have mentioned in Section 3.7, there is a lot of research ongoing on the credibility of information and information sources, and we do not address this topic in our work. In a future version of our model, we consider to incorporate the main ideas on this subject derived from the computational reputation research field.

4.2.1 Context of Agreements

We have mentioned in Section 2.2.1 (Proposition 1) that trust is situational: it depends on the context in which it is measured and it depends on the

particular situation within this context. Hence, a computational trust model shall estimate the trust on a trustee in the particular situation and context under assessment, instead of deriving a situation-less global measure of trust. Before we present our proposal for a situation-aware computational model of trust, we need to clarify the notion of context used in this work.

In our work, we restrict the analysis of context to dyadic interactions between a truster and a trustee where the latter must perform a task on behalf of the former. Both the truster and the trustee may be individuals or organizations. We define context in order to characterize the situation of past agreements of the trustee under assessment and to characterize the present situation in which a trust decision is needed. By knowing the trustworthy behavior of the trustee in different situations in the past, we may undercover tendencies of behavior of the trustee that are situational and reason about them in the current situation.

Following (Abowd et al., 1999), context is any information that can be used to characterize the situation of an entity. This work further considers that there are four main types of context: identity, location, activity, and time. In the same way, a system that uses contextual information as part of its operation is then considered context aware (Strang et al., 2003).

In our model, we draw on Abowd et al. (1999)'s work to consider that context is expressed by eight dimensions $d_1, d_2, ..., d_8$.

$$\underbrace{d_1 \quad d_2}_{\text{identity}} \quad \underbrace{d_3}_{\text{time}} \quad \underbrace{d_4}_{\text{location}} \quad \underbrace{d_5 \quad d_6 \quad d_7 \quad d_8}_{\text{activity}}$$

This way, dimensions d_1 and d_2 correspond to the identity context type and identify the truster and the trustee of the reported interaction, respectively; d_3 represents the time of the agreement; d_4 relates to the location context type; and d_5, d_6, d_7, and d_8 identify and characterize the type of the task, its complexity, deadline, and the outcome of its realization, respectively. The use of all these dimensions is not mandatory in all contexts, with the exception of dimensions d_1, d_2, d_5 that must be present in all contexts, and d_8 that must be present when representing an individual item of evidence concerning any given past interaction. In certain contexts, additional dimensions may be required to allow higher levels of expressiveness in describing the complexity of a task, and therefore dimension d_6 may be further divided into $d_6^{'}$, $d_6^{''}$, etc.

Our representation of context involves the following steps:

1. Identify the relevant types of dimensions that best characterize the situation. Dimensions d_1, d_2 and d_5 are mandatory.
2. Identify the need to subdivide dimension d_6.
3. Define the set of possible values for each dimension.

4.2.2 Outcome of Agreements

In this work, each task $t_i \in \mathcal{T}$ has a finite number of possible outcomes. Most of the time, we consider that the set of possible outcomes of a given interaction is given by $\mathcal{O} = \{F, Fd, V\}$, where the outcome F means that the truster perceives the agreement to be totally fulfilled by the trustee, in terms of deadline and quality; the outcome Fd means that the truster perceives the trustee to introduce a delay in the realization of the task, although the final result is of good quality; and the outcome V means that the truster perceives a severe violation from the trustee (e.g., the task was not performed, the delay was excessive, or the quality was way below acceptable).

Other levels of expressiveness could have been used to represent the truster's subjective evaluation of how the trustee fulfilled his obligations in performing the task at hand. For instance, in Equation 4.1, we add additional detail allowing to explicitly represent the fact that the deadline was fulfilled but with perceived poor quality (outcome Fpq), or that the truster perceived a fulfillment with a delay and poor quality (outcome $Fdpq$).

$$\mathcal{O} = \{F, Fd, Fpq, Fdpq, V\} \ . \tag{4.1}$$

Regardless of the level of expressiveness used, the reporting outcome is a subjective evaluation of the trustee's behavior that may be sensible to the situation at hand. For example, if the trustee accomplished the assigned task five seconds after the task's deadline, some trusters may perceive the task as fulfilled (outcome F) and others as fulfilled with a delay (outcome Fd). If the task is not critical, it is probable that the vast majority of trusters do not bother signaling a violation of deadline.

Also, the semantic of the outcomes is stable around situations, but the preferences of agents over the outcomes may depend on the context being considered. For example, John may prefer that Mary arrives half an hour late (outcome Fd) over her bringing beer instead of wine (which would be sensed by John as a poor quality outcome, Fpq), whereas an importer of fabric may prefer to receive less product (what he could represent as a 'not so good' outcome, Fpq) over receiving all product with a big delay (outcome Fd) if store replenishment is in danger.

The relative preference relation over the possible outcomes depends then on the agent, the matter, and the specific situation at hand. A logic (but not unique) relation of preferences over the set $\mathcal{O}$ shown in Equation 4.1 is given in Equation 4.2, where $o_1 \succ o_2$ means that outcome o_1 is strictly preferred to outcome o_2, following the theory of rational agents.

$$F \succ Fd \succ Fpq \succ Fdpq \succ V \ . \tag{4.2}$$

4.2.3 Application of Our Contextual Representation

In this subsection, we look at different examples that illustrate the adequacy of our representation of context to practical scenarios.

First Example. Mary reports her knowledge about John's cooking skills in two different events: "John prepared a dinner after Christmas but the food was not good" and "John redeemed himself six months later preparing a delicious snack". These reports could be gathered by the truster and systematized as follows:

Mary	John	jan 2012	cook	high	*Fpq*
Mary	John	july 2012	cook	low	*F*
d1	**d2**	**d3**	**d5**	**d6**	**d8**

In this example, dimensions d_4 (location) and d_7 (deadline) were considered irrelevant to describe the context of this type of agreement. In contrast, the complexity of the task is relevant, as an individual may be able to cook simple meals but fail to cook more elaborate dishes. In the current version of our model, we consider that the degree of complexity of the task is explicitly reported by the interacting partner that evaluated the agreement (in the example, Mary). This may introduce some undesirable noise in some cases; for example, it is possible that Mary is not aware of the complexity involved in both meals – in certain cases, preparing a snack may demand higher cooking skills and time than preparing a dinner – and reports a opinion that may or may not be shared by John. An alternate solution would be to explicitly refer the type of meal (e.g., appetizer, snack, main course) in dimension d_6, possibly using an ontology allowing for share understanding among the participants in a system.[1]

In the first event of the example above, John cooked a dinner, hence fulfilling the agreement of cooking for Mary, but the outcome of this agreement

[1]Such an approach is addressed in (Strang et al., 2003; Tavakolifard et al., 2008).

was of poor quality (outcome *Fpq*). In the second event, John prepared a snack that was delicious and fulfilled the agreement (outcome *F*).

Second Example. An importer of textile fabric is evaluating a given exporter named T1 (the trustee), based on reports provided by agents E1 and E2, which have previously interacted with the trustee.

E1	T1	july 2012	India	cotton	high	high	*Fd*
E2	T1	july 2012	China	denim	low	low	*Fpq*
d1	**d2**	**d3**	**d4**	**d5**	**d6**	**d7**	**d8**

In this case, all eight dimensions are important to describe the importer/exporter context.

Third Example. This example is adapted from (Tavakolifard et al., 2008), and considers two reports of Alice about Bob's ability as a tour guide, in two different situations: Trondheim at night, and Bergen when it is stormy.

Alice	Bob	Trondheim	guide	medium	F
Alice	Bob	Bergen	guide	high	F
d1	**d2**	**d4**	**d5**	**d6**	**d8**

In this case, we omit the time and deadline dimensions. In the same way, we assume that a tour guide at night is of medium complexity, and that this complexity raises to high when it is stormy. Alternatively, the set of possible values for the complexity of the tour guide's task (dimension d_6) could have been set to *night*, *stormy*, and the like (in Tavakolifard et al. (2008)'s ontology-based approach, *night* is a time *aspect*, and *stormy* is a new aspect named *weather*).

Fourth Example. This example is adapted from (Rehák et al., 2008), where humanitarian aid entities in rescue missions need to make trust decisions about transporters of cargo of different degrees of sensitivity (e.g., medical supplies, food and durable goods), delivering in different quantities through roads with varied quality. In the first event, transporter $T2$ transported medical supplies (high sensitivity, dimensions d_6) in high quantities ($d_6^{'}$) through low quality roads ($d_6^{''}$). In the second event, the same transporter transported durable goods (low sensitivity) in high quantities through roads with medium quality.

In this example, the specificity of the rescue scenarios implies high expressiveness in the description of the task, and therefore dimension d_6 was

E3	T2	transport	high	high	low	F
E3	T2	transport	low	high	medium	F
d1	**d2**	**d5**	**d6**	**d6'**	**d6"**	**d8**

divided in sub-dimensions. The authors did not mention the role of time and location in this context, so we did not instantiate dimensions d_3 and d_4.

Fifth Example. This example is adapted from (Nguyên and Camp, 2008), in the domain of selection of services from nodes in *ad hoc* networks. These authors consider that context comprises a number of attributes such as the identity of the client (d_1), the identity of the server (d_2), the identity of the service (d_5), and the date of the interaction (d_3). They also consider other context attributes, such as the reason to invoke the service, the location of client and server (i.e., the number of hops between the two, d_4), the standard deviation of shadowing (σ, d_6), and the path-loss exponent (η, d'_6). In the example, we do not instantiate the reason to invoke the service, as the authors do not make explicit the use of this context attribute.

C	S	12/12/2012	4	S*	8	5	F
d1	**d2**	**d3**	**d4**	**d5**	**d6**	**d6'**	**d8**

4.2.4 Current Situation and Past Evidence

In the sequence of our characterization of context, we represent any situation $s_i \in \mathcal{S}$ as a tuple of values ascribed to each contextual dimension but the last (corresponding to the outcome dimension), as shown in Equation 4.3. If the situation under assessment is the current situation, the value corresponding to the time dimension (i.e., d_3) may also be omitted. In the equation, $v_j^{s_i}$ is the value ascribed to dimension j in situation s_i.

$$s_i = \langle v_1^{s_i}, v_2^{s_i}, ..., v_7^{s_i} \rangle \ . \tag{4.3}$$

An individual item of evidence e_i is also represented using a tuple of values ascribed to each contextual dimension, but now the outcome dimension d_8 is mandatory, as shown in Equation 4.4.

$$e_i = \langle v_1^{e_i}, v_2^{e_i}, ..., v_8^{e_i} \rangle \ . \tag{4.4}$$

As we have already mentioned, a dyadic interaction resulting from an agreement may generate two individual items of evidence. For instance, in the first example of the previous section, the agreement established between

John and Mary could have included an obligation saying that Mary should bring a bottle of wine, and therefore her behavior as a guest could have been evaluated by John as well. In this case, two individual items of evidence would have been produced: e_1, where $v_1^{e_1} = john$ and $v_2^{e_1} = mary$, and e_2, where $v_1^{e_2} = mary$ and $v_2^{e_2} = john$. As it become evident, both evidence items would also report different values for d_5, d_6, and d_8.

As we have seen before, $\mathcal{E}$ represents all evidence available on all agents of society $\mathcal{A}$. The set of all items of evidence existing about a given trustee y is given by $E_{*,y} = \{e_i \in \mathcal{E} : v_2^{e_i} = y\}$. This representation is generic in the way that it can be used in centralized trust services, where all evidence is provided to agents by a unique entity, or decentralized trust services, where the agents are responsible for acquiring this evidence, for example, by maintaining a memory of the past experiences with other agents of the society and by querying other agents. Therefore, $E_{x,y}$ represents all evidence about the direct past experiences of agent x with y, such that $E_{x,y} = \{e_i \in \mathcal{E} : v_1^{e_i} = x, v_2^{e_i} = y\}$. Finally, our representation also allows to represent the evidence on a given agent y that is made available to a given agent z; for instance, $E_{x,y}^z \subseteq E_{x,y}$ is the set of all evidence on agent y that resulted from his interaction with agent x that is held by agent z.

4.3 The SOLUM Framework

The SOLUM framework is composed of different components that globally implement the trust function of social-based agents, taking into consideration the properties of trust that we postulated in Chapter 2.

Figure 4.1 provides an overview of how these components fit together in a global computational trust framework. In this figure, the rectangular elements represent the main evaluation functions considered in the framework and the cylindric elements represent the evidential data sets. We can also observe that different parameters feed the different evaluation functions: the *current situation* ($s \in \mathcal{S}$) represents the situation under assessment, as defined in Section 4.2.4; the *perceived kinship* ($k \in [0, 1]$) indicates how close and related the trustee is to the truster, as estimated by the latter based on the physical and cultural characteristics manifested by the former; the *reputation* ($r \in [0, 1]$) includes reputation-like information about some specific aspect of the trustee's trustworthiness that is not considered in set $E_{*,y}$; the *truster's disposition* ($d \in [0, 1]$) captures the truster propensity to trust and the truster *emotional state* ($m \in [0, 1]$), referred to in Proposition 6.

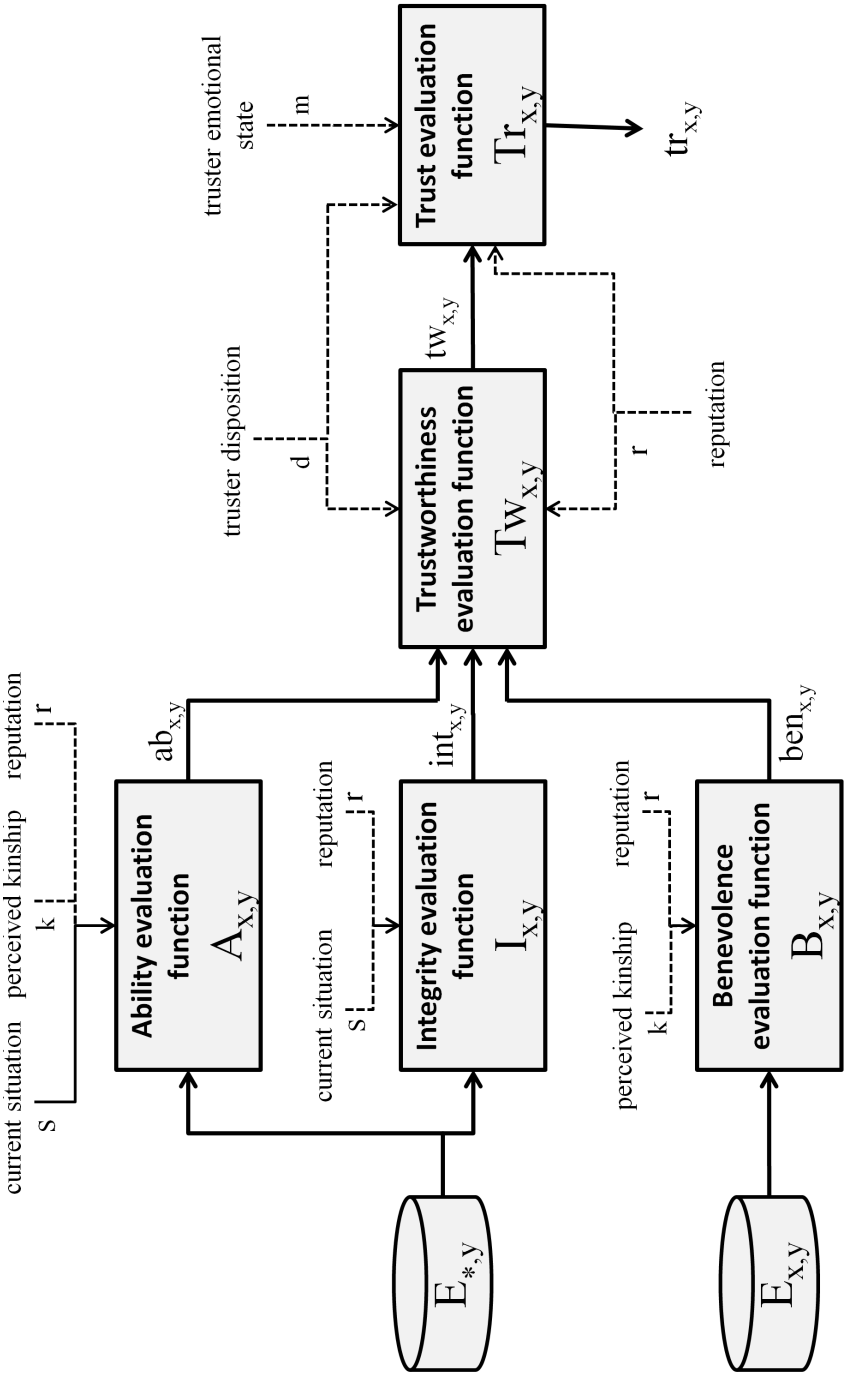

Figure 4.1: The SOLUM framework.

Finally, we used solid lines to represent the functionalities of the framework that we instantiated into computational components, that we present later in this chapter. In turn, elements in dashed lines are not yet implemented.

Using this framework, an agent playing the role of truster estimates the trust it put on a given trustee, in a given situation (Proposition 1). This estimation is mainly a cognitive process involving the estimation of the trustee's trustworthiness (function $Tw_{x,y}$), although the framework also considers the emotional content of trust inherent to the emotional state of the truster (m) and his disposition, or propensity to trust (d) (Proposition 2).

As we mentioned in Section 2.3.5, reputation (r) may also be considered an antecedent to trust (Proposition 25). In the current version of the SOLUM framework, we consider that parameters r, d, and m are produced by some functions outputting values in $[0, 1]$. Hence, the evaluation function $Tr_{x,y}$ is responsible for building an estimate of the truster's trust on the trustee from the estimated value of this trustee's trustworthiness, the emotional state of the truster, his disposition to trust, and (possibly) the reputation-like information about the trustee, in accordance to Proposition 6. Finally, the estimated trust score $tr_{x,y}$ returned by this function is a value in $[0, 1]$, it is not a yes (I trust) or no (I do not trust), following Proposition 5. In fact, although not represented in the figure, we consider that the resulting estimated trust value $tr_{x,y}$ is intended to be used as another element of the decision process of the truster agent. In this case, it may happen that in some situation a truster estimates his trust on a trustee to be 0.8 and still he does not act on this trust, where in other situation he may decide to rely or to be dependent on a trustee for who he estimated a trust value of 0.7. Therefore, our framework does not model the behavioral content of trust (Proposition 4).

We have been arguing since the first chapter of this thesis that the trustworthiness of an agent is multi-dimensional construct that captures the trustee's competence and character (Proposition 7). Contrary to most of the existent computational trust approaches, which consider just one dimension of trustworthiness, we propose and define distinct evaluation functions for the three trustworthiness dimensions considered in (Mayer et al., 1995): ability, integrity and benevolence. We believe that other trustworthiness dimensions appearing in the literature, such as predictability, may be derived from Mayer et al. (1995)'s three basic dimensions.

Next, we describe with a little more detail each one of the evaluation functions that make part of the SOLUM architecture. Later, we present our

proposal for the instantiation of each of these functions.

4.3.1 The Ability Evaluation Function

The ability evaluation function $A_{x,y}$ estimates the general competence of the trustee under evaluation in performing a given task t in a specific situation s, taking into consideration propositions 8 and 9. This function takes as input the evidence available on the trustee under evaluation, $E_{*,y}$ (as defined in Section 4.2), the situation s under evaluation; the perceived kinship k, and the specific (mostly reputational) information r that may exist relating to the specific ability of the trustee, which is not included in set $E_{*,y}$. The output of the ability evaluation function is the estimated ability of the agent, $ab_{x,y}$, defined in $[0, 1]$. Function $A_{x,y}$ is represented as shown in Equation 4.5.

$$A_{x,y}: \quad E_{*,y} \times \mathcal{S} \times [0,1] \times [0,1] \longrightarrow [0,1] \ . \tag{4.5}$$

Concerning our approach to instantiate this function, we propose to use any aggregation engine that is able to produce trustworthiness scores from *all* the available evidence on the trustee ($E_{*,y}$) to estimate his ability. We consider that the use of a large amount of data about the trustee is a coarse-grained good indicator of both his very good or very bad ability, in all situations. One possible aggregation engine to be used is *Sinalpha*, a heuristic-based model that we have developed in an early stage of this thesis' work that takes into consideration the asymmetry and perseverance principles of trust (cf. propositions 31 and 32, respectively). We describe *Sinalpha* in Section 4.4.

However, more interesting than using a general aggregator is to be able to perform situation-aware estimations of the ability of trustees. Instead of adapting one of the few proposals of situation-aware computational trust that we reviewed in Section 3.5, which are based on similarity-based measures and/or trust ontologies, we developed a new component from scratch, *Contextual Fitness*. This component, which we present in Section 4.5, is able to extract the tendencies of behavior of the trustee under evaluation in different situations from the evidence available on that trustee, and does not require complex configuration or the predefinition of similarity measures. Moreover, it was designed to produce usable results even when the available evidence is scarce (cf. Research Question 2) and to be modular enough to be used with other components of the framework.

Figure 4.2 illustrates the way in that the outcome of the *Contextual Fitness* component is used to tune the general estimation computed by the

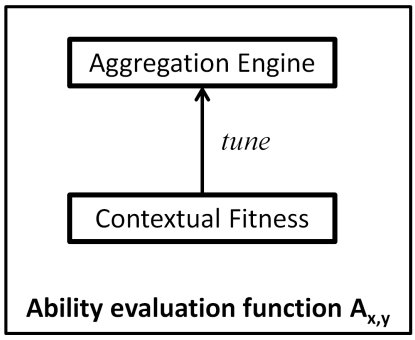

Figure 4.2: Current instantiation of the ability evaluation function, using *Contextual Fitness*.

aggregation engine (e.g., *Sinalpha*) in order to transform it into a situation-aware estimation of the ability of the trustee under evaluation.

4.3.2 The Integrity Evaluation Function

The integrity evaluation function $I_{x,y}$ estimates the integrity of the trustee under evaluation from the set of all evidence available on that trustee, $E_{*,y}$, the situation under assessment, s, and specific (reputation-like) information about the general integrity of trustee, r. The estimated value of integrity $int_{x,y}$ outputted by this function reflects the consistency of the trustee's actions in accordance with the principles established with his interacting partners. Following Proposition 16, a trustee that acts with integrity in one sphere of the individual's life does not necessarily assure that he is going to act with integrity in another sphere of life, although integrity is generally much less situational than, for example, ability. Nevertheless, we consider to input the situation in the integrity evaluation function, as referred before.

We represent function $I_{x,y}$ as shown in Equation 4.6.

$$I_{x,y}: \quad E_{*,y} \times \mathcal{S} \times [0,1] \longrightarrow [0,1] \ . \tag{4.6}$$

In order to implement function $I_{x,y}$, we developed the *Integrity Tuner* component, which checks for the consistency of the behavior of the trustee and the compliance to the social norms implicit in the agreements he established with others (cf. Proposition 17). *Integrity Tuner* is described in Section 4.6.

4.3.3 The Benevolence Evaluation Function

The benevolence evaluation function $B_{x,y}$ estimates the benevolence of the trustee under evaluation toward the truster from the evidence set $E_{x,y}$ corresponding to the direct experience between both agents. Additionally, this function also receives as input the perceived kinship (k), in accordance to Proposition 12, and specific (reputation-like) information (r) about the general benevolence of the trustee with others. The estimated value of benevolence $ben_{x,y}$ outputted by this function reflects different benevolence-based factors, such as the affect that the trustee has developed toward the truster, which commonly arises in long-term and close relationships (Property 13), the trustee's disposition to benevolence that is related with his traits of personality (Proposition 11), and the satisfaction with the relationship (Proposition 14), which leads to positive mutualistic benevolence (Proposition 15).

All these factors make benevolence a dyadic construct and that is the reason why function $B_{x,y}$ is input with the evidence base $E_{x,y}$ and not $E_{*,y}$. We represent function $B_{x,y}$ as shown in Equation 4.7.

$$B_{x,y}: \quad E_{x,y} \times [0,1] \times [0,1] \longrightarrow [0,1] \ . \tag{4.7}$$

Concerning our proposal to instantiate function $B_{x,y}$, we developed the *Social Tuner* component, which estimates the trustee's specific attachment toward the truster and his disposition to do good to the truster, from the evidential set $E_{x,y}$. The *Social Tuner* component is described with detail in Section 4.7.

4.3.4 The Trustworthiness Evaluation Function

The trustworthiness evaluation function $Tw_{x,y}$ estimates the trustworthiness of trustees from the individual estimates of ability ($ab_{x,y}$), benevolence ($ben_{x,y}$), and integrity ($int_{x,y}$). According to Proposition 20, the disposition of the truster (d) is also input into function $Tw_{x,y}$. In the same way, reputation information is considered by some authors as also influencing the estimation of trustworthiness (Proposition 25), and therefore it is considered an input to $Tw_{x,y}$.

This function outputs the estimated score of the trustee's trustworthiness $tw_{x,y}$, which is a value in [0,1]. Equation 4.8 represents this function.

$$Tw_{x,y} : [0,1] \times [0,1] \times [0,1] \times [0,1] \times [0,1] \longrightarrow [0,1] \ . \tag{4.8}$$

Although the SOLUM framework does not impose any particular solution for the instantiation of this function, we propose different ways of

combining the estimated values of ability, integrity, and benevolence, as described in Section 4.8.

4.3.5 The Trust Evaluation Function

The trust evaluation function $Tr_{x,y}$ estimates the trust that a given truster has on the trustee under evaluation. Following Proposition 6, this function takes as input the estimated value of the trustee's trustworthiness ($tw_{x,y}$), the truster's disposition to trust (d), reputation information (r), and the emotional state of the truster (m). The effect of this last input in trust is further supported by propositions 2 and 3. This function, whose representation is shown in Equation 4.9, returns the estimated value of the trust that truster x has in trustee y, $tr_{x,y}$.

$$Tr_{x,y} : [0,1] \times [0,1] \times [0,1] \times [0,1] \longrightarrow [0,1] \ . \tag{4.9}$$

Having described each one of the evaluation functions that compose the SOLUM framework, we are in conditions of describing our proposal to instantiate these functions, which we do throughout the remaining of this chapter.

4.4 The *Sinalpha* Component

Sinalpha is an aggregation engine that takes as input all evidence on the trustee under evaluation, $E_{*,y}$, and computes an estimation of this trustee's general ability, in $[0,1]$, as formalized in Equation 4.10. [2]

$$Sinalpha : E_{*,y} \rightarrow [0,1] \tag{4.10}$$

Sinalpha was first designed as a general trustworthiness estimator following properties of the dynamics of trust (namely, the asymmetry and perseverance of trust), in an early stage of our work. Later, it was adopted as one possible implementation of the situation-less component of the ability evaluation function. In this way, it is intended to provide general estimations of any trustee's ability or competence, irrespective of the particular situation and of the specific relationship that might exist between this trustee and his evaluator. A high to very high value of the estimated ability construct indicates that the trustee is generally competent, and a low to very low value of this score indicates that the trustee is rather incompetent or

[2] *Sinalpha* first appeared in (Urbano et al., 2009a). A detailed description of its genesis in given in (Urbano et al., 2012).

has not shown his trustworthiness yet (e.g., because he is a newcomer). Intermediary values may be rather uninformative and a deeper understanding of the trustee's other qualities, such as his integrity and benevolence toward the current truster, would be necessary in order to make a correct estimation of the trustee's trustworthiness.

The algorithm of *Sinalpha* entails two main steps:

1. Update parameter α from the set of all evidence on the trustee.
2. Compute the situation-less estimate of ability, $sinalpha(\alpha)$.

Step 1 – Update α. We use parameter *alpha* to aggregate the behavior of the trustee manifested through the outcomes of his agreements with others throughout his lifetime, as known by the truster. When the trustee is a newcomer, the truster assigns him an initial value α_0 which may reflect either ignorance or *prior* knowledge about the trustee. For example, if the truster perceives that the trustee has specific qualities that enable him for the task at hands (cf. Proposition 9), α_0 can be made higher than when the trustee is a total stranger to the truster. After this first assessment by the truster, the value of α is updated every time a new piece of evidence about the trustee is aggregated. Equation 4.11 presents the update function of α.

$$\begin{aligned} alpha: \quad & [-\pi/2, \pi/2] \rightarrow [-\pi/2, \pi/2] \\ & \alpha_{t+1} = \alpha_t + \lambda \times \omega \ . \end{aligned} \tag{4.11}$$

In Equation 4.11, we use indexes on parameter α to refer to a moment in time $t+1$ that is posterior to moment t; we do not use these indexes elsewhere in this chapter. The value of α grows with the aggregation of evidence related to positive behavior of the trustee, and decays with evidence related to negative behavior, where this growth/decay behavior is controlled by parameter λ. Also, how much α grows or decays with every new piece of evidence is controlled by parameter ω. We describe parameters λ and ω with more detail right after describing Step 2.

Step 2 – Compute $sinalpha(\alpha)$. After parameter α is updated to reflect new knowledge about the trustee, the situation-less general estimation of the trustee's competence is computed using the *sinalpha* function, as defined in Equation 4.12.

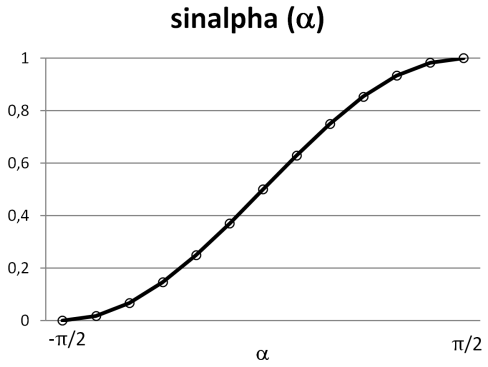

Figure 4.3: The aggregation function of the *Sinalpha* component.

$$\begin{aligned} sinalpha: \quad & [-\pi/2, \pi/2] \rightarrow [0,1] \\ & sinalpha(\alpha) = p \times (\sin\alpha + 1)\ . \end{aligned} \tag{4.12}$$

Function *sinalpha* is a monotonic aggregation function with a sinusoidal shape (Figure 4.3). As shown in Equation 4.12, it is a sine function restricted to domain $[-\pi/2, \pi/2]$ and parametrized by constant $p = 0.5$, in order to restrict the codomain to $[0, 1]$. It is easily verifiable that the lowest and highest values of the trustee's perceived competence are $sinalpha(-\pi/2) = 0$ and $sinalpha(\pi/2) = 1$, respectively. As we have mentioned before, the truster may want to distinguish between ignorance about the trustee's qualities and distrust about these qualities, by setting α_0 (corresponding to the lack of evidence on the trustee) to a value somewhat higher than $-\pi/2$, which is the value that corresponds to total distrust.[3] However, the model does not impose any restriction concerning the value of α_0.

As *Sinalpha* was first designed as a trustworthiness estimator and unique predictor of trust, its function embeds the asymmetry principle referred to in Proposition 31 (cf. Section 2.5) through parameter λ (cf. Equation 4.11). In fact, this parameter assumes positive values (λ_+) when aggregating outcomes associated to positive events, and negative values (λ_-) when aggregating outcomes from negative events. By setting $|\lambda_+| < |\lambda_-|$, we guarantee that negative events have stronger impact on decreasing trust than positive events on increasing trust.

Moreover, we allow the absolute value of λ for negative events to in-

[3]In systems that are not able to handle attacks related to identity changes, it is not advisable that α_0 is much greater than $-\pi/2$. An analysis of different types of attacks to trust systems is given in Kerschbaum et al. (2006).

crease as more negative events are reported, allowing to better penalizing intermittent behaviors. For this, we define the *lambda coefficient*, ρ_λ, as a function of the outcomes o^{e_i} of the agreements established by the trustee so far, as shown in Equation 4.13. Function lf can be instantiated for a given set $\mathcal{O} = \{F, Fd, V\}$ by making $lf(F) = 0$, $lf(Fd) = 0.5$, and $lf(V) = 1.0$.

$$\rho_\lambda = \sum_{i=1}^{|E_{x,y}|} lf(o^{e_i}) \ . \tag{4.13}$$

Then, the value of λ is updated using ρ_λ, as shown in Equation 4.14. In this equation, we assume again that $\mathcal{O} = \{F, Fd, V\}$. Possible values for λ_F, λ_{Fd} and λ_V are 1.0, -0.5, and -2.0, respectively.

$$\lambda = \begin{cases} \lambda_F & \text{if } o^{e_i} \text{ is } F \\ \lambda_{Fd} \times (e^{\rho_\lambda}/100 + 1) & \text{if } o^{e_i} \text{ is } Fd \\ \lambda_V \times (e^{\rho_\lambda}/100 + 1) & \text{if } o^{e_i} \text{ is } V \end{cases} \tag{4.14}$$

On the other hand, parameter ω of Equation 4.11 permits to set how fast or how slow an agent can be perceived as highly trustworthy. In Figure 4.3, ω was set to have value $\pi/12$, meaning that a trustee that is a complete stranger to the truster can be consider by this as fully trustworthy after presenting twelve positive events in a row (cf. the limit property referred to in Jonker and Treur, 1999). Accordingly, ω can be configured differently according to the personality of the trusting agent (e.g. more or less cautious) or even to the specificity and severity of the situation under assessment.

The sinusoidal shape of the Sinalpha function allows to implement the perseverance of trust, as referred to in Proposition 32. In fact, we observe in Figure 4.3 that when the truster perceives the trustee's trustworthiness to be either very high or very low – both cases contemplate the presence of the emotional content of trust–, the curve of the function is flatter than in other stages of the truster's trust, which allows for the truster to stick with his previous convictions even in the presence of (sporadic) contradictory information.

4.4.1 Final Remarks About Sinalpha

We started designing *Sinalpha* following the intuition that the paths of trustworthiness grow and decay of a given trustee should not be the same, but instead describe a route similar to a hysteresis curve. Interesting enough, we came across the conceptual model of Straker that we described in Sec-

tion 2.3.6. Despite the simplicity of this model and the fact that the hysteresis shape was not used to capture our first intuition[4], it addressed in some form the time dimension of relational trust, assuming that the trustee has to walk a path of trustfulness before he can be considered trustworthy, roughly corresponding to Propositions 28 to 30. Moreover, the flat shapes of the hysteresis curve when the trustee is perceived as either very untrustworthy or very trustworthy suggested compliance with the principle of perseverance that we annunciated in Proposition 32. Hence, our first idea was to use the non-linear hysteresis shape in our aggregation engine. For that, we used Lapshin (1995)'s formula depicted in Equation 4.15, where a represents the coersitivity parameter, m and n are integers used to fit the curve, and b_x and b_y are the saturation parameters. Figure 2.7 was obtained making $a = 0.1$, $b_x = 0.5$, $b_y = 0.5$, $n = 3$ and $m = 1$.

$$\begin{aligned} x(\alpha) &= a\cos^m\alpha + b_x\sin^n\alpha, \\ y(\alpha) &= b_y\sin\alpha. \end{aligned} \tag{4.15}$$

We prototyped and tested this idea (cf. Urbano et al., 2009b) against the weighting means by recency algorithm of the FIRE aggregation engine (Huynh et al., 2006). However, we were not able to observe the expected benefits. For instance, we observed that the 'Taking Advantage' phase (cf. Figure 2.7) was too smooth and allowed for severe deceptive behaviors from agents that have proved to be trustworthy in the past. Other limitations of the model were already described in Chapter 2. Given all these considerations, we reconsidered our approach in light of new knowledge about social trust and adapted it to the current version of *Sinalpha*.

As we mentioned before, *Sinalpha* was first designed as a trustworthiness estimator. Soon after realizing that trustworthiness estimation should account for different dimensions of trustworthiness, such as ability, benevolence, and integrity, we stopped further testing *Sinalpha*, as we realized that the real contribution to trust estimation would come from a different type of exploration of the available evidence and less from a deep tuning of the aggregation engine. Therefore, we have the notion that we could have further tunned *Sinalpha*, for instance, trying different values of λ and ω or setting a time window to explicitly implement forgiveness, but that the improvement that we could get from this tunning would not be as relevant as the one we could get by extracting information about the benevolence and

[4]In fact, later on we adapted this intuition to the asymmetry principle of Slovic (1993) (cf. Proposition 31).

integrity of the trustee under evaluation, or even exploring the situation under assessment.

4.5 The *Contextual Fitness* Component

The second component of the ability evaluation function is *Contextual Fitness*. This component takes as input all evidence existing on the trustee and extracts the most probable behavior of the trustee for each possible contextual situation. Hence, it relates the past behavior of trustees with particular situations, allowing to estimate the 'fitness' of the trustee to the situation in assessment. This componenent is intended to work in conjunction with any aggregation engine that output values in $[0, 1]$, as is the case of *Sinalpha*. Other engines can be used (e.g., some of the ones reviewed in Chapter 3) provided that an adequate transposition of the range of output values is previously performed.

In *Contextual Fitness*, the situation is characterized using the definition of context given in Section 4.2.1. More specifically, dimensions d_5, d_6, d_7 and d_8 are expressive enough to allow *Contextual Fitness* to infer that some agents may have more difficulty in agreements involving highly complex tasks, while others may default in the presence of tight deadlines, or any other conjugation of these two dimensions.[5] The algorithm for *Contextual Fitness* entails two main steps:

Step 1 – Prepare the evidence. In the first step of the algorithm, a dataset corresponding to dimensions d_5, d_6, d_7 and d_8 is created from all evidence that the truster has on the trustee under evaluation, $E_{*,y}$. The possible values for each dimension are chosen: v_5 includes all possible values identifying the task (e.g., *cook*, *organize event*), v_6 and v_7 include all values for the complexity of the task (e.g., $v_6 = \{low, medium, high\}$) and deadline (e.g., $v_7 = \{small, medium, big\}$), respectively; and v_8 has values in a subset of $\mathcal{O}$ (e.g., $v_8 = \{F, Fd, V\}$, where F stands for fulfillment, Fd represents a contingency (e.g., fulfillment with a delay), and V represents a violation of the agreement). The resulting dataset is then a set of tuples $\langle v_5^i, v_6^i, v_7^i, v_8^i \rangle$, denoted by $E'_{*,y}$.

[5]The dimensions of context are chosen manually by the human, and ideally they appear frequently in the set of past evidence about the trustee. Some examples in different domains are: "The supplier delivered three containers of material in two weeks, as agreed", "The student returned the books in good condition, but with three days of delay", and "Mary returned the call as soon as she could".

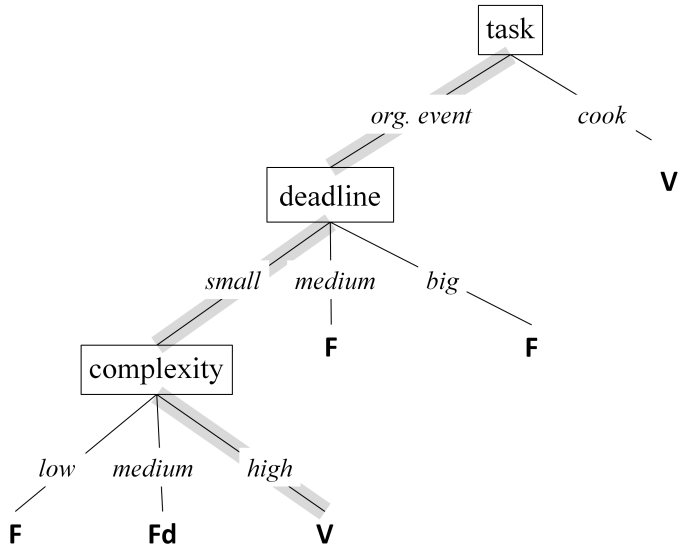

Figure 4.4: Classification tree encoding the evidential set of a trustee.

Step 2 – Infer the trustee's fitness to current situation. In this step, the information gain-based ID3 classification algorithm (Quinlan, 1986) is applied to dataset $E'_{*,y}$, with the attribute corresponding to d_8 settle as class attribute. In the resulting learned tree, each node represents one contextual dimension d_i, the edges that leave out the node are the possible values v_i considered for this dimension, and the leafs correspond to possible behaviors represented by outcomes o_i. Therefore, this tree gives information about the most probable outcomes to be delivered by the trustee in the situations that are described using the contextual dimensions considered to build the tree.

As an illustrative example, Figure 4.4 shows the learned tree that classifies the evidential instances of a trustee in two different tasks: organization of events and cook, with each task being characterized by its complexity (with values in $\{low, medium, high\}$) and deadline (with values in $\{small, medium, big\}$).

In order to infer the fitness of the trustee to a given situation, we need to classify the instance that represents the situation. For instance, if we consider the situation under assessment as represented by instance $s = \langle task = org.\ event, complexity = high, deadline = small \rangle$, then moving down the tree branches corresponding to the values of task, deadline, and complexity (highlighted in gray, in the figure), we get that this trustee has a tendency to produce outcome V (i.e., to violate the agreement) in this situation. In the same way, we observe that this trustee has a tendency

to violate any agreement that involves cooking, irrespective the complexity and deadline associated to the task.

To resume, *Contextual Fitness* receives as input the set of evidence on the trustee under evaluation ($E_{*,y}$) and the representation of the situation under assessment (s) and outputs the most probable outcome o_y^s for this trustee and this situation, as represented in Equation 4.16.

$$\textit{Contextual Fitness} : E'_{*,y} \times \mathcal{S} \rightarrow \mathcal{O} \ . \tag{4.16}$$

It is important to refer that the algorithm just described is repeated every time the truster needs to estimate the trustworthiness of the trustee, which means that we are not using the ID3 algorithm as an offline classification process, with separate training and testing phases, but instead as an online and incremental process. This brings three different types of benefits. First, it allows *Contextual Fitness* to work even when the dataset available on the trustee under evaluation is very small, which allows us to answer positively to part of the research question 2 that we annunciated in Section 1.2.1. Second, as the estimated tendency of behavior of the trustee in a given situation changes dynamically with the size of the historical data on this agent, *Contextual Fitness* turns to be very responsive to any change in the trustee's behavior. And third, contrary to other situation-aware approaches that rely on predefining similarity distance functions and/or ontologies, our model is incremental and does not require other predefined information than the choice of the contextual dimensions and possible values inherent to evidence representation. Section 5.3 presents the experimental analysis of *Contextual Fitness* and the results obtained, which confirm the three benefits just mentioned.

4.5.1 Application of *Contextual Fitness*

Contextual Fitness is meant to be used with other computational trust-based components. For instance, it can be considered a computational trust add-on allowing situationless trustworthiness estimators (e.g., the ones used in the Beta Reputation reputation function, cf. Equations 3.1 – 3.2, or in the aggregation function of FIRE presented in Section 3.3) to turn into situation-aware models. In the same way, we envision different possible ways in which the output of this algorithm (i.e., the most probable outcome in the situation under assessment) can be conjugated with a trustworthiness score; for example, the estimated trustworthiness score may be set to zero if the trustee shows a tendency to violate his agreements in current situation; or the extracted tendency may be used as a discount factor of the esti-

mated score of trustworthiness. An evaluation of both methods in different simulated scenarios is presented in Section 5.3.

In computational trust approaches that privilege the notion of social trust, as is the case of SOLUM, another question arises: which of the trustworthiness dimensions shall be considered contextual? From Proposition 8, we know that ability is domain and task-specific, and then it makes all sense to use *Contextual Fitness* when estimating the ability of a trustee; on the contrary, propositions 11 – 13 make no claim about benevolence being contextual: this is a construct that expresses a feeling of goodwill of the trustee toward the truster that is not sensitive to context. Finally, the literature on integrity is not clear about the effects of context, but, as stated in Proposition 16, it is possible that integrity slightly changes outside specific spheres of life. In this thesis, we use *Contextual Fitness* when estimating the ability of partners, as described in Section 4.8, and we leave for future work the use of this component when estimating the integrity of the trustees.

4.5.2 Final Remarks about Contextual Fitness

The first version of *Contextual Fitness* used a version of the Frequency Increase metric (Paliouras et al., 1999) to extract the situation-based tendencies of behavior of trustees (Urbano et al., 2011d). The algorithm proved efficient when the trustees had some kind of handicap in one context dimension, but has shown some difficulties in revealing handicaps in more than one of these dimensions (e.g., the trustee showing a tendency to fail when the task is of high complexity *and* is due in a low period of time). The first version of *Contextual Fitness* using the Information Gain metric was first published in (Urbano et al., 2010b).

4.6 The *Integrity Tuner* Component

The *Integrity Tuner* component is our proposal to instantiate the integrity evaluation function $I_{x,y}$ defined in Section 4.3.2. In our current proposal, we do not implement the use of specific reputation-like information, nor the (possible) situation-awareness of the integrity estimation. The challenge we faced was to extract any information available about the trustee's integrity from the evidential set $E_{*,y}$, as represented in Equation 4.17.

$$\textit{Integrity Tuner} : E_{*,y} \rightarrow [0, 1] \ . \tag{4.17}$$

Being aware of the difficulties inherent to modeling a construct as complex as integrity from a structured and reduced set of evidence, we still

believe that the inference of the consistency of the trustee's actions and their accordance to the principles shared with his partners may help tuning the estimation of the trustee's trustworthiness. In this respect, we hypothesized that the use of such information would improve the reliability of the trustworthiness estimation (Hypothesis 3).

Particularly, we focused on one particular aspect of the integrity definition as postulated in propositions 16 and 17: a person of integrity is a person that remains true to his relevant social commitments. This means that we look for the trustees' tendency to be consistent with their actions (and then to produce consistent outcomes) and to fulfill their promises. Hence, our instantiation of the *Integrity Tuner* component is based on two complementary coefficients: the *coefficient of consistency*, which captures the trustees' consistency in the delivered outcomes; and the *coefficient of promises fulfilled*, which measures how well the trustees fulfill their promises. We describe both coefficients in the next subsections.

4.6.1 Coefficient of Consistency

The *coefficient of consistency* captures the degree of heterogeneity or 'disorder' in the past behavior of the specific trustee under assessment. We know from information theory that Shannon entropy is a measure of disorder or uncertainty of a probabilistic system (Shannon, 2001), and therefore it seems an elegant solution to our purposes.

This way, if we consider that the outcomes of the past transactions with the trustee under assessment follow a distribution over values in $\mathcal{O}$ (considering, for example, $o_1 = F$, $o_2 = Fd$, $o_3 = V$), then we can define the entropy of past outcomes as given in Equation 4.18, where $p(o_i)$ is the probability mass function of outcome o_i and n is the cardinality of $\mathcal{O}$.

$$H(\mathcal{O}) = -\sum_{i=1}^{n} p(o_i) \times \log_b p(o_i) \ . \tag{4.18}$$

We further normalize the value of entropy, by dividing it by the maximum value of entropy, which is $\log_b n$. As we use normalized values of entropy, the choice of b is not relevant. We consider $b = 10$, although $b = 2$ and $b = e$ would yield the same results.

Finally, we define the *coefficient of consistency*, $\rho_{cs} \in [0, 1]$, as given in Equation 4.19. A trustee that is consistent in his deliverables produces a low value of normalized entropy, and then ρ_{cs} is high, and the other way around.

$$\rho_{cs} = 1 - \frac{H(\mathcal{O})}{\log_{10} n} \ . \tag{4.19}$$

4.6.2 Coefficient of Promises Fulfilled

In the previous subsection, we described the *coefficient of consistency*, which measures the consistency of the trustee's past actions. However, the consistency of a trustee does not guarantee his integrity, as the trustee can consistently perform badly. We need to make sure that the trustee is consistent in fulfilling his promises. That is, that he performs the task within the time frame agreed upon and that the outcome of the task is of good quality. In this sense, it is evident that in a normal condition a trustee that fulfilled a given agreement in its plenitude (outcome F) acted with integrity, whereas the trustee that completely violated the agreement (outcome V), for the same conditions, *may* have acted with no integrity.

Above, we referred to a normal condition, meaning that the environment was not deterring the agent to perform the task. For instance, if John was caught in a situation of urgency and could not call Mary to cancel the dinner, he would have violated the agreement of cooking for her but he would not necessarily be acting without integrity. In the same way, the textile exporter that fails to deliver three containers of cotton due to an unforeseen dock strike is not necessarily acting without integrity. Nevertheless, we use the number of total or partial fulfilled outcomes to tune the estimated value of the trustee's integrity, as calculated by the *coefficient of consistency*. This number is given by the *coefficient of promises fulfilled*, $\rho_{pf} \in [0, 1]$, calculated as shown in Equation 4.20.

$$\rho_{pf} = \max(\frac{\sum_{i=1}^{N_y} vlr(o^{e_i})}{N_y}, 0) \ . \tag{4.20}$$

In the equation above, N_y is the number of all agreements established by trustee y in the past, i.e., $|E_{*,y}|$; and vlr is a function $vlr: \ \mathcal{O} \rightarrow [\Re_0^-, 1.0]$ that indicates how much the truster values each possible outcome for his agreements. Considering $\mathcal{O} = \{F, Fd, V\}$, possible values for this function are $vlr(F) = 1.0$, $vlr(Fd) = 0.5$, and $vlr(V) = -0.5$. Using these values, we consider that outcome F is twice as desirable in terms of promises fulfilled as outcome Fd, and that outcome of type V is not desirable at all.

4.6.3 Estimating the Trustee's Integrity

Taking into consideration the two coefficients calculated previously, we are in conditions of presenting our formula to estimate the integrity of trustee y, $int_{x,y} \in [0, 1]$, as shown in Equation 4.21.

$$int_{x,y} = \begin{cases} \rho_{pf} & \text{if } N_y = 1 \\ \rho_{pf} & \text{if } N_y > 1 \text{ and } \rho_{pf} < 0.5 \\ \frac{1}{2}\rho_{cs} + \frac{1}{2}\rho_{pf} & \text{otherwise} \end{cases} . \tag{4.21}$$

The above formula considers that it does not make sense to evaluate the consistency of the trustee's actions when there is only one evidence reporting on these actions. In the same way, it is not relevant to consider the consistency coefficient when the trustee is consistent in his bad behavior, producing a low value for ρ_{pf}. In this respect, we fixed the value of 0.5 empirically, assuming that $\mathcal{O} = \{F, Fd, V\}$ and that $vlr(F) = 1.0$, $vlr(Fd) = 0.5$, and $vlr(V) = -0.5$. Other value should be used in different conditions. Finally, in all other situations, the estimated integrity of the trustee under evaluation is given by the simple mean of the values of ρ_{cs} and ρ_{pf}.

4.6.4 Final Remarks about *Integrity Tuner*

As we have mentioned before, the current version of the *Integrity Tuner* component aims at improving the reliability of the estimated value of trustworthiness of any given trustee y, using information from $E_{*,y}$. This does not mean, however, that the use of other types of information about the integrity of y (e.g., specific reports or reputation on y's integrity) is worthless, but that these other sources of information may be unavailable and then we have to rely on the information extracted from $E_{*,y}$. Hence, we envision the use of *Integrity Tuner* as complementary to other components that extract trustworthiness-related information from $E_{x,y}$, such as *Sinalpha* or *Social Tuner*, a component that we present in the next section.

In the same way, we understand the estimated value of integrity as calculated by *Integrity Tuner* as an asset to be used wisely when making trust judgments, and not necessarily as a number to be careless summed or multiplied by in some formula. For this reason, we complemented our exploratory and analytical work on integrity with an experimental, simulated-based study where we tried different uses of the *Integrity Tuner* component in the SOLUM computational trust model. We concluded from that study that

there are particular circumstances that denounce that a given trustee has low integrity, despite presenting relative high values of global trustworthiness; we also concluded that in trust-based scenarios of selection of partners, we are able to improve the trust judgments and consequently the selection decision if we exclude these low-integrity agents from selection. We present the evaluation of *Integrity Tuner* in Section 5.5.

4.7 The *Social Tuner* Component

The *Social Tuner* component is our proposal to instantiate the benevolence evaluation function $B_{x,y}$ defined in Section 4.3.3. In our current proposal, we do not implement the use of the perception of kinship neither the use of specific reputation-like information, and therefore the challenge that we faced was to extract any information available about the trustee's benevolence toward the truster from the evidential set $E_{x,y}$. In this respect, we hypothesized that the use of such information would improve the reliability of the trustworthiness estimation (Hypothesis 1).

However, how to model a construct as complex as benevolence from the set of existing evidence on past interactions between the truster and the trustee, even more when the used representation of evidence is necessarily structured and not very verbose (cf. Section 4.2.4)? We realize then that any approach to benevolence in such conditions could not be comprehensive in covering the benevolence concept. However, we believe that our initial purpose of getting more from the available set of evidence by extracting benevolence-like information from this set, in order to increase the reliability of the estimated trustworthiness, still maintains its validity. Equation 4.22 represents function *Social Tuner*.

$$\textit{Social Tuner} : E_{x,y} \rightarrow [0,1] \ . \tag{4.22}$$

This way, the *Social Tuner* component measures the trustee's specific attachment toward the truster, i.e., and his disposition to do good to the truster. This is done using the *coefficient of benevolent actions* parameter, which we present in the next subsection.

4.7.1 Coefficient of Benevolent Actions

The *coefficient of benevolent actions*, $\rho_{ba} \in [0,1]$, measures the trend of contingencies presented by the trustee to the truster in the last Δt period of time. These contingencies may be felt by the truster as more or less severe. For instance, when considering $\mathcal{O} = \{F, Fd, V\}$, it is possible that the truster

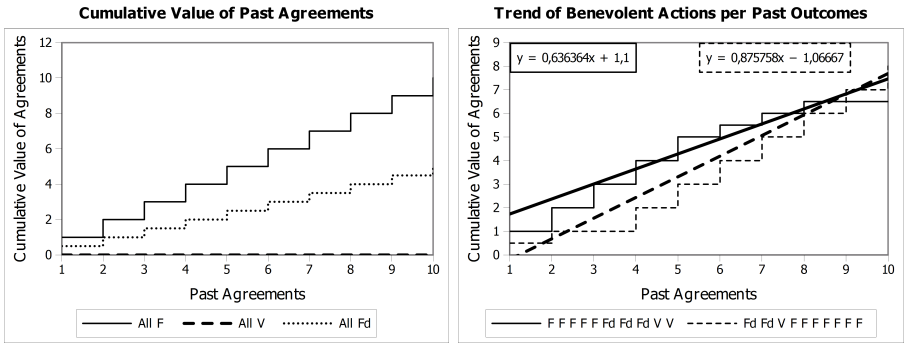

Figure 4.5: Examples of the cumulative outcomes values function (*left*) and of the benevolent actions per past agreements (*right*).

may consider that the outcome Fd corresponds to a mild contingency, while V would be perceived as a severe contingency (for more about the truster's preferences on possible outcomes, see Section 4.2.2).

Hence, the first step to calculate the trend of contingencies is to define how much the truster values each possible outcome for his agreements, using function vl, as represented in Equation 4.23. In the examples presented in this section using $\mathcal{O} = \{F, Fd, V\}$, we consider that $vl(F) = 1.0$, $vl(Fd) = 0.5$, and $vl(V) = 0.0$.

$$vl: \quad \mathcal{O} \rightarrow [0, 1] \ . \tag{4.23}$$

Then, we build a function of the cumulative value of past agreements per generated outcome, $cumValAgreem$, as shown in Equation 4.24. Figure 4.5 (*left*) illustrates the cumulative values of of agreements for three different trustees, each one having interacted 10 times with a given truster in the past, where one of them fulfilled all agreements with the truster, the other delayed all the agreements, and the remaining violated all agreements.

$$cumValAgreem(i) = \sum_{j=1}^{i} vl(o^{e_j}) \ . \tag{4.24}$$

Finally, the *coefficient of benevolent actions* is given by the correlation between the number of agreements established between truster and trustee in the past and the function of the cumulative value of past agreements calculated for these agents. In order to get this correlation, we apply a linear regression to the function of cumulative value of agreements. Figure 4.5 (*right*) illustrates this process for two different agents: one that is very observant of his obligations toward the truster in the first agreements estab-

lished between them but that inverted this behavior in the last agreements, and the other presenting an opposite behavior.

Equation 4.25 reminds the linear regression function for one predictor. In our case, it is used to indicate the progress of the cumulative value of the past agreements, where X represents the past agreements and Y the cumulative values.

$$Y = B_0 + B_1 X \ . \tag{4.25}$$

We use the intercept (B_0) and the regression coefficient (B_1) to estimate if the trustee's benevolence toward the truster is steady, is progressing positively, is progressing negatively, etc. This means that by using this process we are able to estimate how the benevolence of this relationship is evolving. Finally, the coefficient of benevolent actions is given by a function of the correlation coefficient and the intercept, as illustrated in Equation 4.26.

$$\rho_{ba} = B_1 + 0.10 B_0 \ . \tag{4.26}$$

The value of this coefficient is minimum ($\rho_{ba} = 0$) when the trustee constantly delivered the worse possible outcomes (i.e., V) in past agreements with the truster indicating that he was acting with no benevolence at all toward the truster. Conversely, the value of this coefficient is maximum when the trustee totally fulfilled all the past agreements with the truster, showing high benevolence toward him.

4.7.2 Estimating the trustee's Benevolence

The estimated value of the benevolence of the trustee toward the truster, $ben_{x,y}$, is derived from the *coefficient of benevolent actions* using the formula in Equation 4.27.

$$ben_{x,y} = \frac{1}{2}\rho_{ba} + \frac{1}{2}\frac{\sum_{i=1}^{N_{x,y}} vl(o^{e_i})}{N_{x,y}} \ . \tag{4.27}$$

In the equation above, $N_{x,y}$ represents the number of past interactions between truster x and trustee y, such that $N_{x,y} = |E_{x,y}|$. It is worth noting that the estimation of benevolence is only possible when there are, at least, two past interactions between the truster and the trustee under evaluation, i.e., $N_{x,y} \geq 2$. In the same way, this estimated value of the benevolence must be updated at every new trustworthiness estimation, as the benevolence of agents may evolve due to the mutualistic satisfaction/dissatisfaction of the trustee with the relationship, which may change with time and context.

Also, as we have mentioned before, long-term relationships may also generate genuine trust affect toward the other partner of the relationship.

4.7.3 Final Remarks about *Social Tuner*

By evaluating the benevolence of the trustee toward the truster, we are able to account for the *emotional content of trust*. For example, let us imagine that *Sinalpha* (or any other single-dimension trustworthiness estimator) derived a low to medium value of trustworthiness for the trustee under evaluation; this might indicate that the trustee is low in ability, integrity, benevolence, or all three. However, if the *Social Tuner* indicates a high value of benevolence of the trustee toward the truster, this may mean that both partners are engaged in a benevolent relationship, and that the truster may expect the trustee to fulfill a future joint agreement.

In a contrasting example, if *Social Tuner* detects a low benevolence level toward the trustee and the general trustworthiness score of the latter is high, it is highly probable that the trustee has high ability in performing the task, but has low benevolence toward the truster. Knowing this information, the truster can either avoid to enter in a future agreement with the trustee, or give the first step to promote goodwill trust by not denouncing a contingency by the trustee. However, if the trustee's trustworthiness is low, this might indicate that the trustee is either very low in ability or very low in benevolence or integrity (or all three cases), which gives a precious clue to the truster that the trustee is possibly not a good partner to establish an agreement with.

In the examples above, we implicitly indicate that the *Social Tuner* must be used in conjunction with, at least, a single-dimension trustworthiness estimator, such as *Sinalpha*. In Section 5.6, we test different algorithms for the combination of these two components.

4.8 Combining Ability, Integrity and Benevolence

In this section, we instantiate the trustworthiness evaluation function, $Tw_{x,y}$, defined in Section 4.3.4. Our current version of this function does not consider the truster's disposition to trust (d) neither the use of reputation (r). What it does is to combine the estimated values of the ability ($ab_{x,y}$), integrity ($int_{x,y}$), and benevolence ($ben_{x,y}$) of the trustee under evaluation, taking into consideration Proposition 18, which states that the benevolence of a given trustee may only be perceived at later stages of the relationship, and Proposition 19, which states that the right balance between the three

Table 4.1: Example dataset.

#	v_1 (truster)	v_6	v_7	v_8	#	v_1 (truster)	v_6	v_7	v_8
1	C2	high	big	F	15	C10	low	big	F
2	C17	low	big	F	16	C20	high	medium	F
3	C6	low	medium	F	17	C7	medium	medium	F
4	C13	high	big	F	18	C2	high	big	F
5	C2	low	low	F	19	C10	low	low	F
6	C11	low	medium	F	20	C1	high	medium	C
7	C8	medium	big	F	21	C5	high	big	F
8	C5	high	low	V	22	C0	high	medium	V
9	C9	low	big	F	23	C15	high	big	F
10	C7	high	low	V	24	C6	low	big	F
11	C3	medium	big	F	25	C3	low	big	F
12	C4	high	medium	F	26	C17	low	big	F
13	C14	medium	low	F	27	C13	medium	medium	F
14	C1	medium	medium	F	28	C20	medium	big	F

trustworthiness dimensions depends on context and developmental phase of the relationship.

This way, by providing the different computational trust components that are able to infer each one of these trustworthiness dimensions, we have the necessary tools to model different day-to-day trust situations. For example, in some situations, integrity is paramount but not benevolence; in other situations, benevolence secure integrity, so integrity is not that important; also, in situations where the partners have never interacted before, the component of benevolence should not be used. This diversity of situations may be modeled by ascribing weights to each one of the dimensions that may be changed in evolving relational conditions. To illustrate this idea, we provide in Table 4.1 a possible evidential dataset for a given trustee; then, in Figure 4.6, we illustrate how the final value of trustworthiness for this trustee as estimated by truster $C7$ can be so different as we consider different weights for each one of the estimated values of the trustworthiness dimensions. This expressiveness in determining trustworthiness scores would not be possible if we were not be able to consider separate values for ability, integrity and benevolence.

As of the writing of this thesis, we are working on different alternate algorithms of function $Tw_{x,y}$ that combine the three trustworthiness dimensions instantiated into $ab_{x,y}$, $int_{x,y}$, and $ben_{x,y}$. We came across two simple approaches that, despite their preliminary stage of development and lack of sophistication, are described in the next two subsections, as possible examples of function $Tw_{x,y}$.

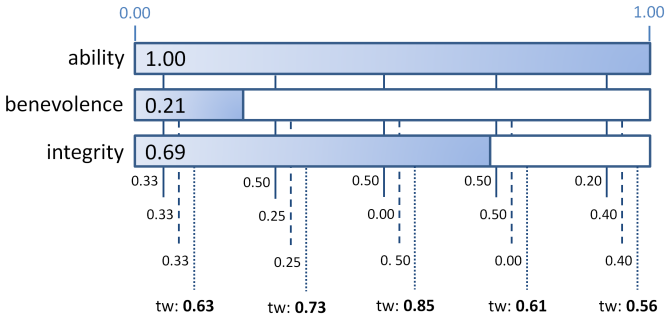

Figure 4.6: Different combinations of ability, benevolence, and integrity.

4.8.1 Function $Tw_{x,y}$ – Alternative One

The first of the two above mentioned approaches to $Tw_{x,y}$ determines the weights that shall be ascribed to each trustworthiness dimension in specific stages of the relationship between truster and trustee, and then performs a weighted mean of the estimated values of $ab_{x,y}$, $int_{x,y}$, and $ben_{x,y}$, using these weights. This approach can be described with the following five-step algorithm:

1 – Determine the weight of benevolence. The weight of benevolence on the trustworthiness score is 0.0 when the truster has just one or none previous interactions with the trustee, and it grows progressively until the truster considers that he has enough experience with the trustee to infer his benevolence toward him. Let $N_{x,y}$ be the number of past interactions between truster x and trustee y and $N_{ben_{min}}$ the minimum number of past interactions between partners that allows benevolence to weight as much as ability and integrity altogether. In the same way, let $N_{ben_{max}}$ be the number of interactions that indicates that the partners are enrolled in a close relationship; this way, when $N_{x,y} > N_{ben_{max}}$, we consider that $N_{x,y} = N_{ben_{max}}$. Then, the weight of benevolence is given by $\omega_{ben} = (N_{x,y}/N_{ben_{min}})/3$.[6]

2 – Determine the weight of integrity. The weight of integrity on the trustworthiness score is 0.0 when the trustee has just one or none previous interactions, and it grows with the number of this trustee's past interactions, N_y. Let we denote by N_{int} the total number of past interactions of the trustee that is considered enough to evaluate his integrity, such that we

[6]This would roughly corresponds to the intimate level of interactions defined in the Regret model (Sabater and Sierra, 2001), where the authors claim that, from a social point of view, when the agents achieve this level they are in the stage of close relation, and more experiences will not increase the reliability of the opinions from then on.

consider that $N_y = N_{int}$ if $N_y > N_{int}$. Then, the weight of integrity is given by $\omega_{int} = (N_y/N_{int})/3$.

3 – Determine the weight of ability. The weight of ability is given by $\omega_{ab} = 1 - (\omega_{ben} + \omega_{int})$.

4 – Penalize ignorance about benevolence. We define the coefficient of ignorance ρ_{ign} that takes value 0.1 when the truster has just one or none previous interactions with the trustee, and 0.0 otherwise.

5 – Estimate the trustworthiness score. Finally, the estimation of the trustee's trustworthiness score is given as shown in Equation 4.28.

$$tw_{x,y} = (\omega_{ab}.ab_{x,y} + \omega_{int}.int_{x,y} + \omega_{ben}.ben_{x,y}) \times (1 - \rho_{ign}) \ . \tag{4.28}$$

4.8.2 Function $Tw_{x,y}$ – Alternative Two

The second approach to $Tw_{x,y}$ separates the use of integrity and benevolence. Although this approach needs further design work, we intend with it to exaggerate the content of Proposition 18, which states that the integrity (and ability) of partners may be perceived earlier in the relationship, and that the perception of these partner's benevolence tend to happen later on the relationship. In order to evaluate such an approach, we propose an algorithm with the following steps:

1 – Define a threshold for benevolence. We define threshold N_{ben}, which represents the minimum number of past interactions between the truster and the trustee under evaluation for considering the benevolence dimension.

2 – Select the $Tw_{x,y}$ function. If the number of interactions between both partners, $N_{x,y}$, is equal or below threshold N_{ben}, the trustee's estimated trustworthiness is given by the combination of the estimated values of his ability and integrity, such that $tw_{x,y} = 1/2ab_{x,y} + 1/2int_{x,y}$. Otherwise (i.e., when partners are enrolled in a long-term relationship), this estimated trustworthiness is given by the combination of the estimated values of the trustee's ability and benevolence, such that $tw_{x,y} = 1/2ab_{x,y} + 1/2ben_{x,y}$.

3 – Use heuristics to further reason about the trustee's trustworthiness. After computing $tw_{x,y}$, the truster may further reason about his *decision* of weather to trust or not trust the trustee, using information-processing strategy heuristics. These heuristics resulted from the insights derived from the literature revision on integrity and benevolence, and from an extensive analysis of the results of multiple experiments we performed with *Integrity Tuner* and *Social Tuner*, individually. We separate this reasoning in the two moments described in step two, as described next.

Hence, when considering the moment described by $N_{x,y} \leq N_{ben}$, we consider another threshold, N_{int}, which is a minimum limit on the number of past interactions of the trustee for considering integrity. When $N_{x,y} < N_{int}$, the truster is not convinced at all of the trustee's integrity if the latter's estimated consistency, or his estimated value of integrity, both calculated using *Integrity Tuner*, are zero. The final trustworthiness of this trustee would be very low, in this case, and, in a scenario of partners' selection, the proposal of this trustee would most probably be removed.[7] When $N_{x,y} \geq N_{int}$, the trustee's integrity is questioned in the same way if his estimated value of integrity is zero or his estimated consistency is lower than any given consistency threshold cst, i.e., when $\rho_{cs} < cst$. We needed this extra step to address the cases when the trustee is very active in establishing agreements, may tend to fulfill most of them, but even so occasionally fails the deadline or violate some of these agreements. In this case, the estimated consistency would be a complementary source of integrity to consider, instead of using only the overall estimated value of integrity.

Finally, when considering the moment described by $N_{x,y} > N_{ben}$, we use heuristics that reflect the global perception of the truster relative to the benevolence level of the population of trustees (the ones for which there is evidence) in his society. This way, the truster periodically updates the mean and maximum values of all trustees' benevolence, and determines the average of these values (let us name it the *mean-max* value of benevolence). Then, he suspects that the benevolence of any given trustee may be less than desirable if this trustee's benevolence is lower than the *mean-max* value. As happened before, the final trustworthiness of this trustee would be very low, in this case, and, in a scenario of partners' selection, the proposal of this trustee would most probably be removed.

[7] In fact, we use the heuristics described here when evaluating the *Integrity Tuner* component individually, in Section 5.5.

4.9 Calculating Trust

In this section, we address the trust evaluation function, $Tr_{x,y}$, defined in Section 4.3.5. In the current version of this function, we do not consider the truster's disposition to trust (d), the use of reputation (r), or the truster's emotional state (m). Consequently, the final trust that the truster has on the trustee, $tr_{x,y}$, is estimated using only the estimated value of the trustee's trustworthiness, $tw_{x,y}$, computed as described in the previous section. This means that, so far it concerns the work of this thesis, $Tr_{x,y} = Tw_{x,y}$.

4.10 Concluding Remarks

In this chapter, we presented the SOLUM model, our agent-based approach to computational social trust, which takes into consideration important properties derived from the theory of trust. We highlighted here three characteristics of trust that are covered by our model but that are often neglected by existing computation trust approaches. First, trust is more than trustworthiness. In this regard, the SOLUM framework considers the existence of other antecedents to trust mentioned in theoretical literature on trust, such as the truster's disposition to trust, the emotional state of the truster, and reputation information about the trustee under evaluation.

Second, trustworthiness is multi-dimensional. As we have seen in Chapter 2, Mayer et al. (1995) reviewed a plethora of work on trustworthiness that propose different antecedents to trustworthiness, and aggregated those antecedents into the ability, integrity, and benevolence dimensions, which we have adopted in the SOLUM framework. However, as we have seen in Chapter 3, only the computational approaches by Castelfranchi and Falcone (2010) and Adali et al. (2011) consider the multi-dimensionality of trustworthiness, and from these only the former presents a computational implementation of such features.

And third, in order to use evidence about the past behavior of a given trustee under evaluation, it is fundamental to understand the benevolence relationships that exist between this trustee and his evaluators, as the outcomes of these past interactions depends on the ability and integrity of the trustee, but also on his benevolence toward the truster.

The SOLUM model is composed of two distinct parts. The first part is the SOLUM framework, a conceptual framework for reasoning about (social-based) trust across a wide range of problems that can be instantiated in different models of trust. This framework is constituted of different functions for the evaluation of the trustee's ability, benevolence, and integrity, as well

as for the evaluation of the trustee's trustworthiness and the trust that the truster has on the trustee. The second part is our current instantiation of the SOLUM framework, which is composed of modular computational components – *Sinalpha*, *Contextual Fitness*, *Social Integrity*, and *Social Tuner* – and a general algorithm to estimate the trustee's trustworthiness from the outcomes of those components, which are combined taking into consideration the stage of the relationship between the truster and the trustee.

All computational components of SOLUM were designed taking into consideration the fact that in open and dynamic agent-based environments, the evidence existing between any truster-trustee pair may be scarce, and newcomers are often a reality. Although different authors remarked the fact that isolated or sparse experiences do not allow to make correct judgments about trustees (Sabater and Sierra, 2001; Burnett, 2011), we also know that in certain scenarios only few – but qualified – experiences are needed to exclude an agent from a community, as we have seen with the example of the diamond trade in New York that we cited in the introductory chapter. With this in mind, we designed our computational components taking into consideration the fact that they should work in a satisfactory way even when the evidence about a given trustee is scarce. This is particularly true with *Contextual Fitness*, which is able to rectify early extraction of situation-aware behavior tendencies through its online and dynamic base algorithm.

To the best of our knowledge, our approach to computational trust is innovative. In fact, it is the first that proposes to extract information about the integrity and benevolence of trustees from the structured evidence existing on these trustees. Also, although the proposal of Castelfranchi and Falcone (2010) already considered different antecedents of trustworthiness, our approach is unique in the sense that it combines these antecedents taking into consideration the stage of the relationship between the interacting partners. For instance, the use of benevolence is less important when these partners are almost new to each other, but is paramount when they have already established an ongoing relationship.

We consider our work a basis that allows to develop complete social-based computational approaches, keeping in mind their immediate adoption in real software systems. Concerning ongoing and future work, we intend to develop more sophisticated instantiations of the SOLUM framework, by considering the unimplemented parts of the model. For instance, we intend to incorporate existing algorithms proposed in the computational reputation research field to implement the use of reputation-like information in the different evaluation functions of the SOLUM framework. In the same way, we are interested in deepening our study on betrayal and relate this with

the truster's emotional state that is input to the trust evaluation function. Also, we intend to further develop each one of the proposed computational components, either conceptually and in terms of computational efficiency. Such as an example, an approach to aggregate old evidence into epochs based on changes of the trustee's behavior, as proposed by Ruohomaa and Kutvonen (2008), would allow for the necessary computational and space management efficiency.

In the next chapter, we evaluate the SOLUM model – more concretely, our current instantiation of the SOLUM framework – through experimental analysis.

Chapter 5

Evaluation of the SOLUM Model

In this chapter, we present the evaluation of our approach to computational (social) trust described in Chapter 4. This evaluation was performed through experiments in an agent-based simulation environment. We conducted two different types of experiments: i) evaluation of the individual computational components instantiating the SOLUM framework, namely, *Sinalpha*, *Contextual Fitness*, *Integrity Tuner*, and *Social Tuner*; ii) evaluation of the integration of all these components, including the relationship-based aggregation of the outcomes of each individual component.

The experiments were conducted at different time spots of the thesis' time, and the requirements of each type of experiments were different; consequently, we developed distinct testbeds with different characteristics for each one of those types. However, all of them shared a common structure, which we present in the next section.

5.1 Introduction

In order to evaluate the contribution of our computational trust approach, we implemented a simulated agent-based system where agents playing the role of trusters sought to select the best trustees to perform specific tasks. The simulation was then performed in a clients-providers' environment. The process of partners' selection included a one-round, multi-attribute negotiation process and resulted in an agreement established between the truster (i.e., the client) that started this process and the selected trustee (i.e., the provider). Although different testbed configurations were chosen for distinct types of experiments, they all shared a common structure with a similar pro-

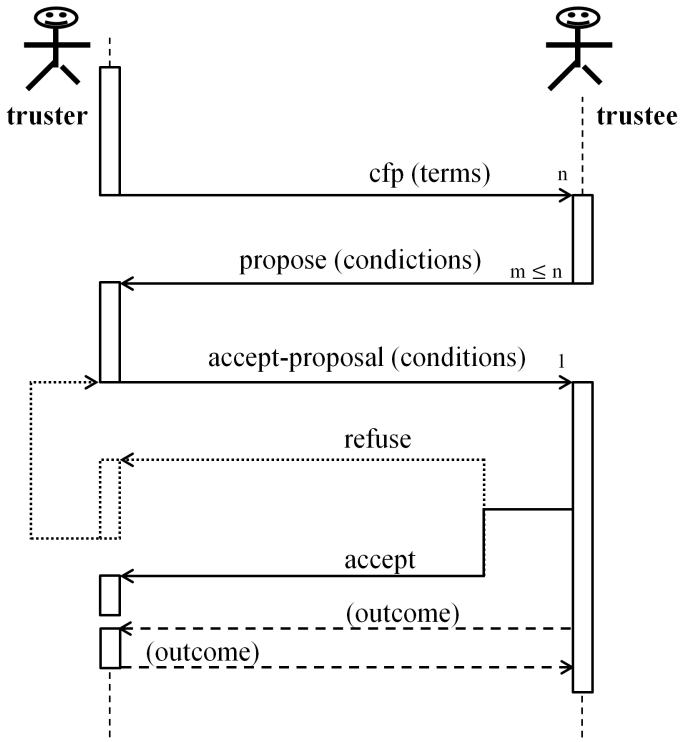

Figure 5.1: Generic process of selection of partners.

cess of selection of partners, which is illustrated in the FIPA AUML-based diagram[1] of Figure 5.1.

5.1.1 Generic Selection Process

Every experiment had a predefined number of rounds, and a different selection process was initiated by each truster at every round. Hence, at each round, every truster identified a given task he wanted to be accomplished by others, as well as the conditions in which he wanted the task to be performed. These negotiation terms were represented using our contextual representation presented in Section 4.2, with the task being represented by dimension d_5, the complexity of the task by dimension d_6, and the deadline by d_7. Additional conditions, such as the price to pay for the service, if considered, were represented by subdividing d_6 (e.g., d'_6 may represent the price). Also, depending on the specific experiment, the values of these dimensions were randomly assigned at setup and fixed throughout all rounds of the experiment, or randomly assigned in the beginning of a new round.

[1] Cf. http://www.auml.org/.

The task and the conditions were then transmitted to all providers of the system (n) through a call for proposals (cfp), as can be observed in Figure 5.1.

The number $m \leq n$ of providers that presented a proposal to the truster in response to the cfp depended on each type of experiment. In some of those types, the providers had a maximum number of proposals they were allowed to respond, in others they had a stock that might or might not prevented them to propose, and in the remaining types of experiments the providers responded to all received cfp. In the same way, in some types of experiments the providers presented their own conditions in the proposals they made (i.e., they presented values for d_6 and d_7 that were different from the ones presented by the client), where in the remaining types the providers just accepted the trusters' terms.

After receiving all proposals, each truster selected the best partner to establish an agreement with based on his own selection criteria. In most of the types of the experiments considered, each truster selected the best rated candidate, which was a decision based on his trust on each one of the the candidates (tr, see Equation 5.1), on the value of each candidate's proposal (up, see Equation 5.2), or on both factors (Equation 5.3). The value of any candidate's proposal, which with a little abuse of the nomenclature we named the utility of the candidate's proposal up, was a function of the proposed values v_6 and v_7.

$$D_x^{tr} = \arg\max_{y_i \in Y} f(tr_{x,y_i}) \ . \tag{5.1}$$

$$D_x^{up} = \arg\max_{y_i \in Y} f(up_{y_i}) \ . \tag{5.2}$$

$$D_x^{tr\_up} = \arg\max_{y_i \in Y} f(tr_{x,y_i}, up_{y_i}) \ . \tag{5.3}$$

In some restricted set of experiments, the highest rated candidate was allowed to not accept to establish an agreement with the truster, even hading previously presented a proposal to this truster. In this case, the truster tried to establish an agreement with the second highest rated candidate, and so one. This option is signaled with dotted lines in Figure 5.1.

Finally, an agreement was established between each truster and the corresponding selected partner. After that, the behavior of the latter would dictate his final attitude toward the agreement, either by fully fulfilling it or by presenting some sort of contingency. The set of all possible outcomes $\mathcal{O}$ that could be assigned to classify the trustee's behavior was defined for each

type of experiment. In some sets of experiments, the behavior of trusters also played a role in defining the final outcome.

5.1.2 Methodology

In all sets of experiments described in this chapter, most of the trusters made use at some point of the estimated trustworthiness of trustees to select partners. In this way, when comparing different computational trust approaches, the truster agents that used the best of these approaches were expected to have the largest number of successful agreements (i.e., those for which the trustee was classified with outcome F) and the smallest number of violated agreements (i.e., those leading to outcome V). Hence, in all sets of experiments, we measured the percentage of each type of outcome belonging to $\mathcal{O}$ obtained by each type of truster agents (e.g., variable `F` measured the percentage of outcomes of type F obtained in a specific experiment). We also measured the percentage of different providers that were selected by each truster type (variable `D`), in order to evaluate how conservative/exploratory each computational trust approach was. In specific sets of experiments, other variables were measured, and they are described in the proper section where they appear.

Every experiment was repeated 30 times and the measured variables were averaged by experiment and truster type. When the difference between the results of some variable in two different models were not evident, we used a Paired Two Sample for Means (one-tail) t-Test to evaluate the statistical significance of the differences. We used Bonferroni adjustments considering the number of t-Test comparisons to be performed in every experiment ($nComp$), where comparisons were considered significant for p-values less than $0.05/nComp$, for experimentwise error rate of 5%.

Finally, each one of the first three main hypotheses that we formulated in Section 4.1 were divided in new (sub) hypotheses that we tested in each set of experiments described in this chapter.

5.2 Evaluation of *Sinalpha*

As mentioned in Section 4.4, *Sinalpha* was our first approach to computational trust. We developed it even before we designed the SOLUM framework, and by this time we were more focused on tuning the algorithm for performance gain than on understanding how trust was affected by complex relations between trusters and trustees. In this early phase of our work, we published different papers describing *Sinalpha* and comparing it

to other trustworthiness estimators, such as the aggregation engine of the FIRE model (Huynh et al., 2006) and the trust update function defined in (Bosse et al., 2007). This work can be seen in (Urbano et al., 2009a; Danek et al., 2010; Urbano et al., 2012).

In this section, we concentrate on testing different values for *Sinalpha* parameters λ and ω. The experiments were conducted in our simulated experimental scenario, where clients chose between different providers having different characteristics of ability and benevolence. For this, we used the benevolence-based model of agents' behavior that we are going to describe in Section 5.4. In all experiments of this section, the process of partners' selection used trust as the only selection criterion, using the $\arg\max_{y_i \in Y} tr_{x,y_i}$ criterion. In all experiments, we ran 60 clients simultaneously and 30 providers. Every experiment had 100 rounds, and was repeated 30 times.

5.2.1 First Set of Experiments

In the first set of experiments, we ran four different types of client populations, each with 15 agents. One of these populations chose the partners randomly, and served as a baseline for comparison purposes. The other populations used *Sinalpha*, each one holding different values for parameters λ_F, λ_{Fd}, and λ_V. More concretely, the first population was configured with $\lambda_F = 1.0$, $\lambda_{Fd} = 0.0$, and $\lambda_V = -1.0$; the second population was configured with $\lambda_F = 1.0$, $\lambda_{Fd} = 0.0$, and $\lambda_V = -1.5$; and the third population was configured with $\lambda_F = 1.0$, $\lambda_{Fd} = -0.5$, and $\lambda_V = -2.0$. In all cases, the value of ω was fixed to $\pi/12$.

Results

The average number of outcomes of type F and V obtained by each population of clients is shown in Figure 5.2 (*top*). We observed that the *Sinalpha*-based population that penalized more outcomes of types Fd and V got 3% more of outcomes of type F and 4% less of outcomes of type V than the other populations, when compared to the baseline population.

5.2.2 Second Set of Experiments

In the second set of experiments, we ran four different types of client populations: one chose the partners randomly (the *baseline* population), and the others used *Sinalpha* with fixed values of $\lambda_F = 1.0$, $\lambda_{Fd} = -0.5$ and $\lambda_V = -2.0$. Hence, the populations differed in the configured value of ω: the first population was configured with $\omega = \pi/6$, which means that any

trustee that was a newcomer could be considered by the truster as fully trustworthy after presenting six outcomes of type F in a row; the second population was configured with $\omega = \pi/12$, needing 12 of these outcomes to be considered fully trustworthy; and the third population was configured with $\omega = \pi/18$ (needing 18 of these outcomes in a row).

Results

Figure 5.2 (*bottom*) shows the average number of different providers selected by each population, which is a good indicator of how much or less the different populations explored new partners. As expected, *Sinalpha* tended to explore more partners when fully trustworthiness was easily achievable ($\omega = \pi/6$) and to explore less partners when more steps were needed to prove the agents' trustworthiness, corresponding to $\omega = \pi/18$. However, the exploration tendency of the three populations was too low (from 0.099 to 0.117) when compared to the exploration tendency of the baseline population (0.798). Consequently, the difference in the results in terms of outcomes of type F, Fd and V was not significant.

Discussion

Although we tried different configurations for *Sinalpha*'s parameters λ and ω, we did not achieve relevant information from these experiments that deserve publication in this thesis. In the same way, we evaluated the penalization of λ with consecutive negative events producing outcomes Fd and V, as described in equations 4.13-4.14, but opted to not show the experiments here, for the reasons described above.

This way, we fixed the values of these parameters in posterior experiments with SOLUM, considering that $\lambda_F = 1.0$, $\lambda_{Fd} = -0.5$, and $\omega = \pi/12$, and we did not further tested *Sinalpha*.

5.3 Evaluation of *Contextual Fitness*

In order to evaluate the contribution of our approach concerning Research Question 2 and to test Hypothesis 2, we ran a set of experiments with *Contextual Fitness*. For this, we settle an environment where the trusters were business clients in the textile industry, and the trustees were providers of textile fabric. In this environment, the context associated to a task in negotiation was given by the fabric to be transacted (dimension d_5), the quantity of this fabric, representing the complexity of the task (dimension d_6), and the deadline for the delivery of the fabric (dimension d_7).

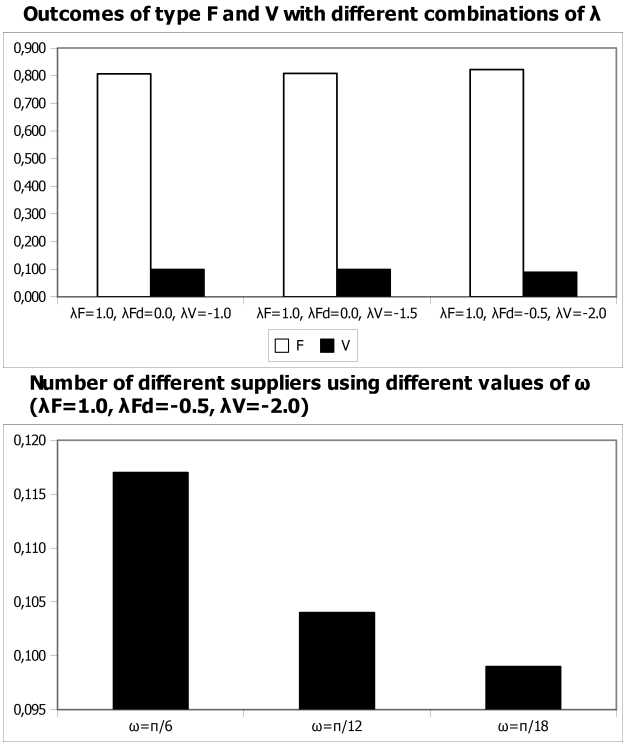

Figure 5.2: Experiments with *Sinalpha*: testing different values of λ and ω.

Then, the providers were modeled to show different handicaps in performing some particular aspect of their tasks. For example, some providers had a tendency to fail to deliver any fabric in short delivery times, while others tended to fail to deliver high quantities of any fabric type, etc. Table 5.1 shows all handicaps that were considered in these experiment. Having a specific handicap means that if a truster selected a provider to transact with in a given situation, and this provider shows a handicap matching the situation, then the provider would fail the agreement (outcome V) with a probability of 95%. Otherwise, this failure probability drops to 5%. As an example, if the truster issues a cfp defining the terms $(cotton, high, low)$ and the selected provider has a handicap in providing high quantities, then he will fail the agreement with a probability of 95%. We assumed that each provider was able to provide all different types of fabric.

In turn, the clients selected the best partners based on their estimated trustworthiness; hence, the aim of these experiments were to evaluate if the use of *Contextual Fitness* increased the performance of the trustworthiness estimator in the presence of populations of trustees behaving differently in distinct situations. In evaluating our approach, we ran different sets

Table 5.1: Handicaps of the populations of providers.

Handicap	Handicap in providing...
`HFB`	a given fabric (chosen randomly at setup)
`HQT`	high quantities of any fabric
`HDT`	any fabric in a short delivery time
`HFBQT`	high quantities of a given fabric
`HFBDT`	a given fabric (randomly chosen) in a short delivery time
`HQTDT`	high quantities in a short delivery time

Table 5.2: Configuration parameters (evaluation of *Contextual Fitness*).

terms of the cfp	randomly assigned at each round
$\mathcal{V}_5$	*cotton*, *chiffon*, *voile*
$\mathcal{V}_6$	*low*, *medium*, *high*
$\mathcal{V}_7$	*low*, *medium*, *big*
$\mathcal{O}$	F, V
terms of proposals	same as in cfp
selection criteria	$D^{tr}_{x,y}$
#clients, #providers	24, 20
#rounds, #exp. repetitions	80, 30
Sinalpha ω, λ_F and λ_V	$\pi/12$, 1.0 and -2.0
Sinalpha $lf(F)$ and $lf(V)$	0 and 1

of experiments, that we describe next. Table 5.2 shows the configuration parameters that are common to all sets of experiments. In the table, $\mathcal{V}_5$, $\mathcal{V}_6$ and $\mathcal{V}_7$ are the sets of all possible values that can be considered for contextual dimensions d_5, d_6 and $d7$, respectively.

5.3.1 First Set of Experiments

In this set of experiments, we wanted to test the following two hypotheses, both derived from Hypothesis 2:

Hypothesis 5 *In the presence of populations of trustees behaving differently in different situations, trusters that are able to extract the behavioral tendencies using* Contextual Fitness *will perform better than those that do not have this ability.*

Hypothesis 6 *The benefits of the* Contextual Fitness *component can be shown when applied to different types of trustworthiness estimators.*

With this intent, we compared the use of our situation-less trustworthiness estimator *Sinalpha* (model `S`) with a trust model consisting of the joint

use of *Sinalpha* and *Contextual Fitness* (model `SC`). Considering that $o_y^s \in \mathcal{O}$ is the estimated tendency of behavior of the trustee under evaluation in current situation s as calculated by *Contextual Fitness*, the trustworthiness evaluation functions of both models are presented in equations 5.4 and 5.5, respectively. As can be observed, the estimated trustworthiness of providers showing a tendency of violating the agreements in the situation under assessment in model `SC` is set to zero, substantially reducing the odds of these providers being selected by the trusters in current selection process. This does not mean, however, that these providers are expected to behave badly in other situations.

$$Tw_{x,y}^{S} = Sinalpha(E_{*,y}) \ . \tag{5.4}$$

$$Tw_{x,y}^{SC} = \begin{cases} Sinalpha(E_{*,y}) & \text{if } o_y^s \text{ is } F \\ 0 & \text{if } o_y^s \text{ is } V \end{cases} \tag{5.5}$$

In the experiments, we ran 12 clients of type `S` and 12 clients of type `SC`. Each one of the 20 providers had a handicap randomly chosen at setup from the values presented in Table 5.1, following an uniform distribution.

Results

Figure 5.3 shows the results of this set of experiments. We verified that, in terms of the number of different providers, `S` ($M = 0.111$, $SD = 0.009$) was less exploratory than `SC` ($M = 0.174$, $SD = 0.017$), $t(29) = -18.88$, $p < 0.05$. This is due to the fact that the agents using *Contextual Fitness* (`SC`) tend to exclude the providers they suspect presenting a handicap in the situation being assessed. We can observe from Figure 5.3 (*bottom, left*) that the more exploratory behavior of *Contextual Fitness* starts as early as the first rounds of the experiments, when the number of evidence available on any given trustee is scarce.

The performance of the two models in terms of successful agreements that resulted in outcome F is shown at the right plots of Figure 5.3. We verified that `S` ($M = 0.808$, $SD = 0.031$) was outperformed by `SC` ($M = 0.859$, $SD = 0.019$), $t(29) = -10.17$, $p < 0.05$, and that this happened since the first rounds of the experiments (Figure 5.3, *bottom, right*). In the conditions of these experiments, we were able to confirm the truthfulness of Hypothesis 5.

In order to test Hypothesis 6, we ran another set of experiments, where

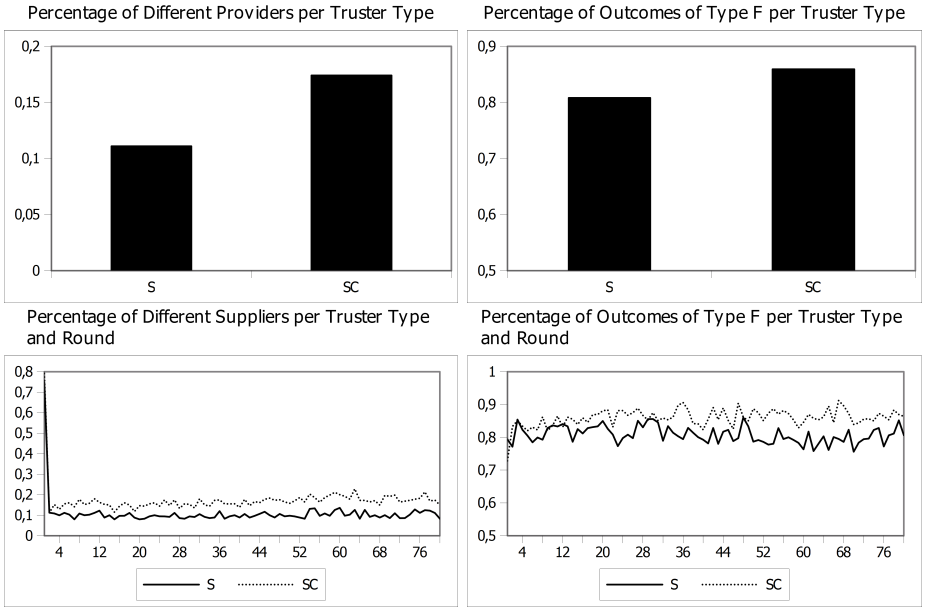

Figure 5.3: Results for *Contextual Fitness* in the first set of experiments.

Contextual Fitness was applied to two other well known trustworthiness estimators. Hence, we ran six different types of trusters simultaneously, each with four agents: `S`, running *Sinalpha*; `SC`, running *Sinalpha* with *Contextual Fitness*; `B`, using the well know Beta Reputation trust aggregation algorithm (Jøsang and Ismail, 2002); `BC`, using the same algorithm along with *Contextual Fitness*; `J`, using the well-known asymmetry-based trust update function defined in (Bosse et al., 2007); and `JC`, using the same algorithm along with *Contextual Fitness*.

The models `B` and `BS` were implemented using equations 3.1 and 3.2 defined in Section 3.3. When aggregating the evidence (Equation 3.2), the the pair $(r^Q_{T,i}, s^Q_{T,i})$ had values $(1, 0)$ for $o^i = F$ and $(0, 1)$ for $o^i = V$. Also, λ was set to 0.9. In turn, the models `J` and `JS` were implemented using equations 3.4 and 3.5 of Section 3.4. In our instantiation of the model, we made $\delta^- = 0.1$ and $e = 0.9$.

The results of these experiments are presented in Figure 5.4. As happened in the previous experiments, the addition of *Contextual Fitness* to models `B` and `J` allowed these models to explore more partners: `S` ($M = 0.288, SD = 0.010$) and `SC` ($M = 0.360, SD = 0.030$), $t(29) = -12.58$, $p < 0.003$; `B` ($M = 0.263, SD = 0.006$) and `BC` ($M = 0.326, SD = 0.028$), $t(29) = -12.46$, $p < 0.003$; and `J` ($M = 0.261, SD = 0.003$) and `JC` ($M = 0.325, SD = 0.034$), $t(29) = -10.22$, $p < 0.003$.

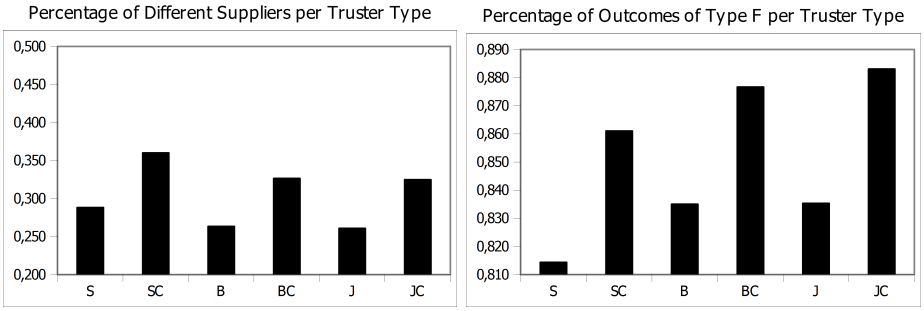

Figure 5.4: Results for *Contextual Fitness* using different trustworthiness estimators.

In the same way, the addition of *Contextual Fitness* allowed all models to significantly increase their performance in terms of well succeeded agreements: `S` ($M = 0.814, SD = 0.026$) and `SC` ($M = 0.861, SD = 0.029$), $t(29) = -7.39$, $p < 0.003$; `B` ($M = 0.835, SD = 0.032$) and `BC` ($M = 0.877, SD = 0.022$), $t(29) = -5.75$, $p < 0.003$; and `J` ($M = 0.835, SD = 0.029$) and `JC` ($M = 0.883, SD = 0.027$), $t(29) = -6.92$, $p < 0.003$. In the conditions of these experiments, we are able to confirm the truthfulness of Hypothesis 6.

Discussion

At a first sight, we could expect that an approach that explores more partners in the scenario described in this section would lead to a smaller number of fulfilled agreements, at least in the first rounds of exploration, where the partners are rather unknown. However, the results obtained in the first set of experiments have shown that the `SC` model using *Contextual Fitness* did not perform worse than the other situation-less approaches in the first rounds and performed significantly better than the other approaches in the remaining rounds of the experiments. This happened due to the ability of *Contextual Fitness* to extract tendencies of behavior even in the presence of a reduced number of available trust evidences, and to its capacity of doing that in an incremental and dynamic way. This allowed that individual bad decisions of trusters concerning the exploration of new providers (which lead to outcomes of type V) were used to update the extracted tendency of failure of the providers in subsequent assessments, approximating the estimated tendency to violate agreements to the true handicap of the providers.

On the contrary, the situation-less approaches tended to select the agents with the highest values of trustworthiness at the moment of the assessment. As handicapped suppliers had the same probability of succeeding outside

the context of their handicap, there was a strong chance that the first partners to be selected by customers were the ones that incrementally increased their trustworthiness. This parochial strategy resulted in the undesirable behavior of trusters keep choosing the same providers that occasionally violated the agreements for which they present a handicap, not giving a chance to explore other providers. In a way, this reflects what succeeds is real life subcontracting in the textile industry.

5.3.2 Second Set of Experiments

In this set of experiments, we wanted to test the following hypothesis:

Hypothesis 7 *In the presence of populations of trustees behaving differently in different situations, trusters that are able to extract the behavioral tendencies using* Contextual Fitness *will perform better than those based on reference contexts regularly placed in a grid.*

With this intent in mind, we compared the use of our situation-aware trust model `SC` with the situation-aware *CSRC* approach (Rehak et al., 2006; Rehák and Pěchouček, 2007; Rehák et al., 2008) that we described in Section 3.5. We used this approach based on predefined similarity metrics to evaluate the benefits of using *Contextual Fitness* when compared to other situation-aware trust approaches.

We name our instantiation of the *CSRC* model as the `RC` model. In this way, we placed the reference contexts regularly over the combinations of all possible values of the contractual attributes.[2] We also chose to compute the trustworthiness of trustees at each reference context r_i by aggregating the new observations using a weighted means by similarity, following Equation 3.6 of Section 3.5. Furthermore, we considered $d(r_i, co)$ as the average of distances $d(r_i, co)^{fabric}$, $d(r_i, co)^{quantity}$ and $d(r_i, co)^{dtime}$, which we describe next. Hence, the distance function we used for attribute fabric is given in Equation 5.6. As can be observed, the distance is minimum (zero) if both contexts c_1 and c_2 have the same fabric and maximum (one) otherwise.

$$d^{fabric}(c_1, c_2) = \begin{cases} 0, & \text{if } fabric_1 = fabric_2, \\ 1, & \text{if } fabric_1 \neq fabric_2. \end{cases} \tag{5.6}$$

[2] As noted in (Rehák et al., 2008), the regular grid configuration is computationally inefficient, as the observations in real applications tend to form clusters instead of spreading uniformly over the context space. However, we used a small context space ($Q = 3$) that did not seriously compromised the computational efficiency.

For the remaining attributes considered in the experiments (quantity and delivery time), the distance function is given in Equation 5.7.

$$d^{attr}(c_1, c_2) = |ln(attr_1) - ln(attr_2)|. \tag{5.7}$$

In the equation above, considering first the attribute *quantity*, $attr_i$ took the value of 1, 3, or 5, depending on the value of the quantity being *low*, *medium* or *high*, respectively. In the same way, for attribute *delivery time*, $attr_i$ took the value of 1, 3 or 5 for values of *low*, *medium* or *big*. The total distance between the two contexts was a weighted means of the three distances calculated above, with all dimensions equally weighted. Finally, the weight used to evaluate the relevance of a context c_1 accordingly to its similarity with context c_2 is given in Equation 5.8. All the remaining formulas needed to compute the trustworthiness scores of agents were implemented accordingly to (Rehak et al., 2006).

$$w_i = e^{-d(c_1,c_2)}. \tag{5.8}$$

In these experiments, 12 agents using our approach to situation-aware computational trust (model `SC`) ran simultaneously with 12 agents using the model `RC` just described. All other configuration parameters were maintained from the previous experiments.

Results

The results obtained are shown in Figure 5.5. From the results, we observed that our approach to situation-aware computational trust, instantiated as model `SC` ($M = 0.175, SD = 0.021$), was a bit more exploratory than the one used in model `RC` ($M = 0.129, SD = 0.017$), $t(29) = 7.46$, $p < 0.05$, although this latter model showed a higher rate of exploration at the initial rounds of the experiment and then progressively decreased the number of different providers explored and stabilized at approximately round 6 (see Figure 5.5, *bottom, left*). In the same way, `SC` outperformed `RC` in terms of successful agreements: `SC` ($M = 0.872, SD = 0.014$), `RC` ($M = 0.841, SD = 0.019$), $t(29) = 7.22$, $p < 0.05$, although this number was approximately equal for both models in the first 6 rounds, the better results of `SC` coming just after that round (see Figure 5.5, *bottom, right*).

We repeated this experiment with a slightly different population of providers, where the set of all possible handicaps were restricted to the values `HFab`, `HQT` and `HDT`. By doing that, we increased the number providers that might present a handicap in the situation under assessment. The results are

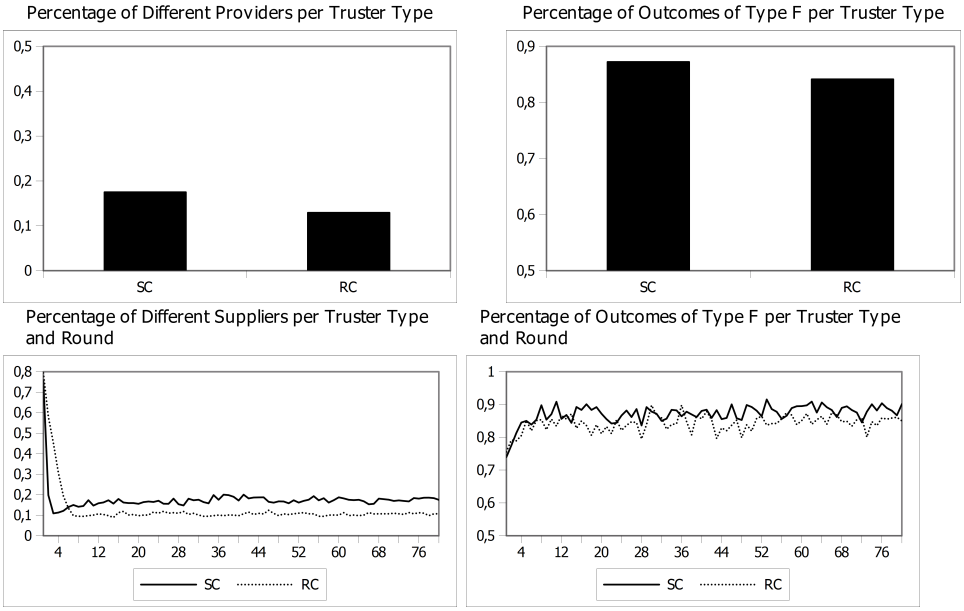

Figure 5.5: Results of the comparison of *Contextual Fitness* with a different situation-aware computational trust approach.

presented in Figure 5.6.

From the results, we observed that `SC` had increased the exploration of different providers per round in about 42% when compared to the previous experiment, due to the fact that there were potentially less providers able to fulfill the conditions of the cfp, and then providers that were known to be trustworthy, as estimated by *Sinalpha*, were probably neglected in comparison with others with lower values trustworthiness but for each *Contextual Fitness* did not estimate a tendency to violate the agreements. This increase was much lower (18%) to `SC`, as this model does not tend to exclude partners based on their tendency to violate the agreements. The final results for this variable are as follows: `SC` ($M = 0.248, SD = 0.013$), `RC` ($M = 0.152, SD = 0.010$), $t(29) = 27.49$, $p < 0.05$.

However, the most significant result for this experiment was in terms of the number of agreements for which a outcome of type F was produced. In both models, this number was reduced when compared to the previous experiment. However, this reduction was much more evident to the `RC` model (about 10%) than to the `SC` model (about 2%), which means that the use of *Contextual Fitness* lead to better estimations of trustworthiness than the use of the *CSRC* model in the presence of strongly handicapped populations. Figure 5.6 (*bottom, right*) shows in a clear way that the `SC` model distantiates from the `RC` from round 4, and that this latter model

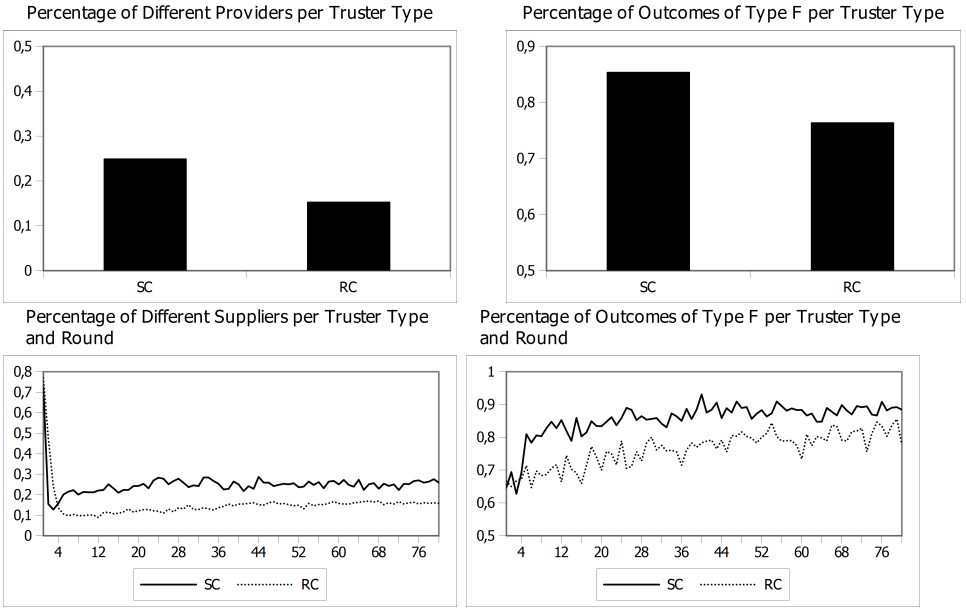

Figure 5.6: Results of the comparison between models SC and RC with providers' populations with single-dimensional handicaps.

needs a great number of information to approximate (yet not reaching) the former's performance. The final results for this variable are as follows: SC ($M = 0.853, SD = 0.019$), RC ($M = 0.763, SD = 0.025$), $t(29) = 17.09$, $p < 0.05$.

In (Urbano et al., 2011c), we ran other experiments comparing the use of *Contextual Fitness* to the *CSRC* model, using different populations of providers, including those that changed their handicap at some point in the rounds of the experiments and those that did not present any handicap. In all these experiments, we observed a better performance of our approach to situation-aware computational trust. Based on all these experiments, we are able to confirm the truthfulness of Hypothesis 7.

Discussion

Concerning the comparison of our situation-aware computational approach to the RC model based on the definition of similarity distance functions, we observed that RC was generally more exploratory than SC in the first four to six rounds of the experiments, after which it started behaving more conservatively, exploring less partners per round. In parallel, RC's performance in terms of fulfilled agreements was poorer than the performance of SC, with particular relevance in the case when the potential number of providers not

Table 5.3: Contractual evidences of a provider (simplified)

evd #	v_5, v_6, v_7	v_8
1	voile, low, medium	F
2	chiffon, low, low	F
3	chiffon, high, medium	V
4	voile, medium, medium	F
5	cotton, low, low	F
6	cotton, medium, big	F
7	voile, low, low	F

showing a handicap in the situation under assessment was lower, where the performance of `RC` was 12% worse than the performance of `SC`. Although not presented in this section, we ran a different set of experiments comparing `SC` with `RC`, where one third of the population of providers shown an erratic behavior, in the sense that their behavior toward fulfillment or violation of agreements was independent of the context of those agreements. In these experiments, which we described and reported the results in (Urbano et al., 2011c), the `RC` model showed to have reduced effectiveness, as the use of reference contexts seemed not to be appropriate when modeling the trustworthiness of these types of mixed populations.

In order to better understand the differences between `RC` and `SC`, we propose to analyze next two examples taken from the experiments we have run.

First example. Table 5.3 shows the set of evidence on a given provider that was generated in one run of the experiments. For the sake of the example, we consider that the terms of the cfp issued by a given truster, defining situation s, is given by $v_5^s = \textit{chiffon}$, $v_6^s = \textit{high}$, $v_7^s = \textit{medium}$, and that this truster is evaluating the proposal made by this provider.

Using `RC`, we verify that only seven of the 27 (3^3) reference contexts defined for the provider <under evaluation directly correspond to the contexts of the past contractual evidences of this provider. We name the reference context corresponding to the situation under assessment as rc_s and we calculate the distance of each individual item of evidence on the provider to this reference context. For example, the first evidence is located at distance $d_{1,s}$ from rc_s, and then rc_s is updated with the value of the evidence's outcome (F) weighted by a function of the distance, as given in Equation 5.8. A similar reasoning is applied when processing the second individual item of evidence. This means that the provider's trustworthiness at rc_s is

increased proportionally to the similarity between the reference context and the context of the two items of evidence.

The third evidence coincides with situation s and therefore the weight w_3 of the evidence is maximum for rc_s, lowering the trustworthiness value at this reference context in a significant way as the outcome of this evidence is F. Finally, the last four items of evidence, all F, raise again the trustworthiness value at rc_s, even if attenuated by the distance of the context of each item of evidence to the context of rc_s. However, due to the fact that the dataset available on the provider under evaluation is very small, the final trustworthiness score for this provider (strongly supported by the rc_s value) is still positive, and therefore bigger than the trustworthiness values of all other providers that have not yet been explored. This explains why, in these conditions, the approach has a tendency to select, from the set of the more fitted providers, the ones that have been involved in more agreements to date, acting in a rather parochial way. From our analysis, we can conclude that the interesting characteristic of bootstrapping of the `RC` shows somewhat disappointing in open and dynamic environments where the available evidence on individual agents can be scarce.

Analyzing now the same example with the `SC` approach, we verify that the *Contextual Fitness* algorithm extracts the tendency of behavior $o_s^t = V$ for this provider in the current situation s (in reality, this algorithm is a bit more expressive in the sense that it is able to detect that this provider tends to violate any agreement that stipulates the delivery of high quantities, for any given type of fabric or deadline), and then the estimated trustworthiness of the provider as calculated by the `SC` model (see Equation 5.5) is zero. Thus, the chance that this provider is selected in current situation is small, allowing the client to select a more adequate proposal or even the exploration of a new partner. We must note that a match between the provider's estimated tendency of failure and the current situation in assessment does not exclude the provider from the selection process, it just lowers it trustworthiness score to zero. In the absence of better alternatives, this provider can still be selected to establish an agreement with the truster.

Second example. Table 5.4 illustrates an excerpt of the evidence set of a provider obtained in another experimental run, using the `RC` approach. The provider under evaluation was configured to be of `HDT` type, which means that he had a 95% probability of violating any agreement specifying short deadlines (i.e., $v_7 = low$). From the table, we observe that the provider was selected several times to establish agreements stipulating low delivery times, indicating that the `RC` model was performing poorly in this specific

Table 5.4: Contractual evidences of supplier *as* (simplified)

evd #	v_5, v_6, v_7	v_8
1	voile, medium, big	F
2	chiffon, high, big	F
3	cotton, low, low	V
4	cotton, medium, big	F
5	chiffon, high, big	F
6	cotton, high, medium	V
7	voile, high, low	V
8	voile, medium, low	V
9	chiffon, medium, big	F
10	voile, low, big	F
11	chiffon, high, big	F
12	voile, medium, big	F
13	chiffon, low, low	V
14	voile, medium, big	F
15	voile, high, low	V

situation.

We found that the problem here concerned the use of the predefined similarity distances among the reference contexts. For instance, let us imagine that a new item of evidence e^{16} was generated such that $v_5^{16} = \mathit{cotton}$, $v_6^{16} = \mathit{low}$, $v_7^{16} = \mathit{medium}$ and $v_8^{16} = F$. Also, let us focus on two specific reference contexts: rc_y, matching situation described by $v_5^y = \mathit{cotton}$, $v_6^y = \mathit{low}$, and $v_7^y = \mathit{low}$, and rc_z, matching situation described by $v_5^z = \mathit{voile}$, $v_6^z = \mathit{low}$, and $v_7^z = \mathit{medium}$. Using the values considered for the `RC` approach in this experimental setting and equations 5.6 and 5.7, we verify that e^{16} is close to reference context rc_y and consequently the trustworthiness of the provider at rc_y would increase in a significant way with the consideration of this new evidence, regardless of the provider's true handicap on low delivery times. On the other hand, the same new item of evidence would increase the reference context rc_z in a less significant way, even though it corresponds to a context for which the provider does not present a handicap. Figure 5.7 illustrates this scenario.

This example shows the limitations associated to situation-aware trust models that rely on predefined measures of similarity. More specifically, an agent that shows a good behavior in a context might fail in what apparently is a *very similar* context and succeed in what appears to be a more *distant* context. Our study of the `RC` approach in the proposed scenario gave us the strong belief that, even dedicating a team of experts to tune the distance

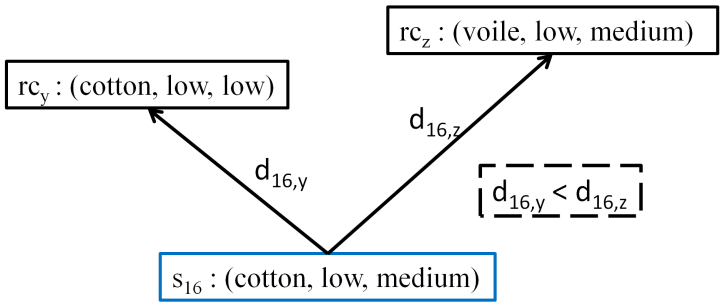

Figure 5.7: Distances between a new evidence and two distinct reference contexts.

functions and/or using taxonomy-based similarity, the use of predefined similarity functions may fail in detecting the contextual subtleties exposed in this last example; in the specific case of the `RC` approach, this hard tuning effort can be even compromised with the addition of more contextual dimensions.

We analyze now the use of the `SC` approach when dealing with the same set of evidence. We proposed that *Contextual Fitness* may be used with extremely small datasets. Hence, we verify that after the third evidence the classification tree build by *Contextual Fitness* returns the following rule corresponding to the V outcome: $v_5 = cotton$. In this case, the algorithm was not be able to detect the true handicap of the provider and would even wrongly lower the probability of this provider being selected to any agreement involving the provision of cotton. Applying the algorithm after the sixth item of evidence, it returns two rules ($r1$ and $r2$) corresponding to outcome V: $v_7^{r1} = low$ and $v_7^{r2} = medium$. This would result in the `SC` client to have a high probability of wrongly missing the opportunity to interact with the provider in agreements stipulating *medium* delivery times. However, the client would also have a high probability of not selecting the provider in agreements involving the delivery in short delivery times (the providers's true handicap), preventing him from making deceitful exchanges.

In this last example, we observed that the `SC` model may be sometimes too restrictive, by overfitting the existing evidence. In a set of experiments reported in (Urbano et al., 2011c), we introduced a population of providers where one third of it consisted of generally bad providers that violated contracts irrespective of the situation under assessment; the other two thirds of the population consisted of providers showing one handicap as described before. In those experiments, we wanted to evaluate if the overfitting characteristic of Contextual Fitness would prevent the `SC` clients from doing good

deals. By adding one third of bad providers, the choice space was reduced in a relevant way and ability of the `SC` in selecting partner fitted to the current situation could be seriously jeopardized due to this overfitting-based generation of failure tendencies. However, the results obtained have shown that the `SC` approach was less penalized than the `RC` approach in this scenario, in line with the results we have shown in Figure 5.6 in a similar situation. Once again, the ability of `SC` to dynamically rectifying the extracted tendencies every time there is a new evidence has shown to be a positive characteristic of the approach.

5.3.3 Third Set of Experiments

In certain areas of world-wide business, business partners choose to adopt parochial environments to the detriment of more aggressive exploration of deals outside the already known partner relationships space (Macy and Sato, 2002). For instance, in the fashion retail industry, clients often rely on knowledge available through textile fairs and textile agents to make the bridge between brands and the reliable textile suppliers. However, even with these guarantees, the space of available suppliers is relatively small and strongly supported by the expected behavior of the partner, rather than on the real utility of the business transaction. The business players would then benefit of computational trust systems that could be used in open and global markets, where the evidence available on the behavior of partners are most certainly scarce, heterogeneous, and contextual.

In this set of experiments, we wanted to evaluate the support given by *Contextual Fitness* to exploring decisions of clients that risk following open market strategies. Therefore, all clients in this particular set of experiments used the `SC` approach. Hence, we wanted to test the following hypothesis:

Hypothesis 8 *In the presence of populations of trustees having different inherent characteristics and behaving differently in different situations, the use of* Contextual Fitness *supports the search for the most desirable partners without jeopardizing the trust-based selection decisions.*

In this last set of experiments, we further divided the clients' population in two: agents of type `Parochial` tended to do business with providers they already know instead of risking new, probably better providers; hence, they selected partners based exclusively on the partners' estimated trustworthiness scores. On the contrary, clients of type `Non Parochial` were aware of the expected value of the business interaction: the selection decision was

based on the providers' estimated trustworthiness and their *internal value* (which we explain next), just that $D_x^{tr-iv} = \arg\max_{y_i \in Y}(tr_{x,y_i} * iv_{y_i})$.

Providers are distinguished by several business individual characteristics, such as their selling prices, international presence, brands they own, quality of their products, the existence of accreditation procedures, organizational strength and economical/financial capacity, and reputation, among other factors (Alves et al., 2012). In this set of experiments, we captured and resumed these different characteristics in parameter iv (internal value); this way, each provider was assigned an internal value that was randomly chosen at setup, following a uniform distribution over values $\{0.50, 0.60, 0.70, 0.80, 0.90\}$. We also assumed that the real internal value of any given provider was only known by a given client after the first interaction between both agents, and that outside these conditions the client would estimate a value of 1.0 for the unknown provider. With this, we wanted `Non Parochial` clients to have the incentive to explore new partners, as their selection decision was based on the partners estimated trustworthiness and internal value. According to what was said before, the selection decision of `Parochial` clients followed Equation 5.1.

In this set of experiments, we measured a new variable, the internal value of the providers selected in each round per truster type, and then calculated the *utility of the agreement*, which took either the value of zero for agreements with outcome V or the internal value of the provider for agreements with outcome F. Additional information about this experiment's configuration is presented in (Urbano et al., 2011c).

Results

Figure 5.8 shows the results obtained per round of experiment. It can be observed from this figure that both approaches got similar results concerning the exploration rate, as given by the percentage of different providers selected by truster and round (*top, left*), and similar results concerning the percentage of agreements with outcome F (*top, right*) . However, the clients of type `Non Parochial` got significant higher values of utility the agreements per round than `Parochial` clients, due to the fact that *Contextual Fitness* allowed them to find the more adequate partners in terms of their contextual fitness to current situation and, at the same time, to find the ones among these with higher internal value. In fact, we can observe in Figure 5.8 (*bottom*) that after the first rounds of exploration, the `Non Parochial` clients got systematically higher utility than the `Parochial` clients.

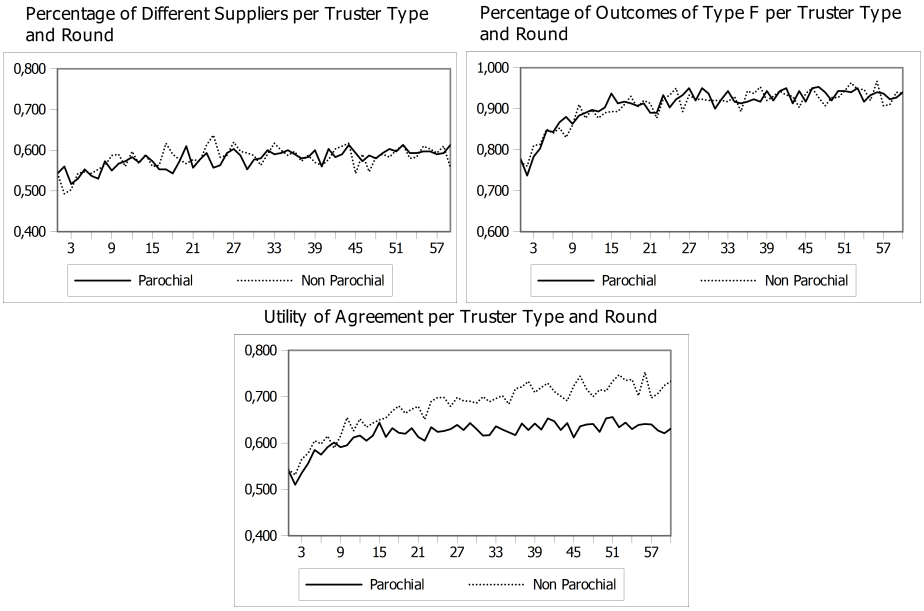

Figure 5.8: Results obtained with `Parochial` and `Non Parochial` consumers.

Discussion

In this set of experiments, part of the clients selected their partners based not only on their estimated trustworthiness but also on their internal value, reflecting the intrinsic characteristics of these partners, which we belief is a more realistic type of decision. The results obtained have shown that the flexibility of the online tendency extraction of *Contextual Fitness* allowed for the clients to explore a larger space of opportunities when searching for partners with more desirable internal characteristics, and yet to do that in a way that does not jeopardize trustworthy choices. Therefore, in the conditions of these experiments, we consider that Hypothesis 8 is true.

5.4 Model of Agents' Behavior

In the previous section, we evaluated *Contextual Fitness* using simple probabilistic models of agents' behavior. However, we believe that, in order to evaluate social-aware approaches of computational trust, which consider the individual dimensions of trustworthiness, and particularly the benevolence dimension, we need more complex models that capture the evolving social behavior of agents. In fact, this is the core idea underlying Research Question 4 introduced in Chapter 1.

For this, we settle an environment where the trusters specified at each round a task with specific requirements of complexity and deadline, and trustees responded by sending a proposal, which could be more close or farther to the trusters' expectations; we measured the value of the proposal p sent by trustee y (up_y) in accordance to the truster's expectations using the variable *utility of the proposal*, `U`. In this environment, we considered that the task under negotiation was the same to all trusters; then, the context associated to this task was given by its complexity (dimension d_6) and the deadline to accomplish it (dimension d_7). In turn, we considered that each truster was characterized by a dispositional benevolence whose value was randomly assigned at setup. In the same way, each trustee was characterized by a value of dispositional benevolence and a value of ability to accomplish the task, both randomly assigned at setup. Moreover, with the dynamics of the negotiations, trusters and trustees were able to develop a form of mutualistic benevolence toward each other. Hence, both trusters and trustees were modeled in a way that covered most of the propositions presented in Section 2.3.1. Thus, they followed a model of behavior principled in several propositions derived from the literature on trust and benevolence, which allowed agents to evolve their behavior based on their interests and on the specific stages of the relationships existing between them. This model is presented in detail in Section 5.4.2.

5.4.1 Motivation

The simulation of social agents in trust-related scenarios is not new. In 1993, Carley et al. (1993) simulated different organizational structures that resulted from the combination of three social characteristics of social agents – honesty, cooperativity, and benevolence – and examined the effect of these behaviors on the cognitive effort, physical effort, communication effort, and idle time, of organizations. However, the social characteristics in this work were defined in a static way. For example, a honest agent always provided correct information, and a cooperative agent always chose to help others before it helped itself. In this work, benevolence concerned the degree to which an agent forgave other agent that provided wrong information.

In other work, Macy and Skvoretz (1998) used a genetic algorithm to test if rules for trusting others could evolve in neighborhood interactions, and if these rules could spread out of the neighborhood through contact with strangers. In this model, the character of the player in exchange, the cultural/physiological and behavioral markers, and the rules for trusting others, were represented in different genes of a chromosome of 15 genes.

Newer approaches to computational trust were evaluated using more simple and static models of agents. In FIRE (Huynh et al., 2006, cf. Section 3.3), one of the most cited models of computational trust, the players on the multi-agent based simulation were consumers and providers of services. Providers were assigned different ranges of competence, defined by a mean and a standard variance. At every round, their performance changed within the range, or even was allowed to switch to a different range. Consumers selected providers by trustworthiness, and their utility was directly connected to the performance of providers. In TRAVOS (Patel, 2006, cf. Section 3.3), providers were assigned probabilities of behaving in a trustworthy/untrustworthy way. The Context Space and Reference Contexts model (Rehak et al., 2006, cf. Section 3.5) was evaluated in a humanitarian aid scenario where providers of transportation services were selected after a major disaster. This selection was based on the providers' trustworthiness and on their bid prices. Although these were based on transportations costs, profit margins, and the providers' competence in specific scenarios, the agents still behaved in a static way.

In (Urbano et al., 2011b), we defined a model inspired in the socio-economic literature where agents decided whether to fulfill their obligations or to present some contingency based on predispositional factors (e.g. benevolence) and on situational factors, such as the importance of the current exchange and whether or not goodwill trust was already formed between the partners. By allowing the behaviors to evolve with time, we have shown that some current computational trust models seem not to be able to fully understand the evidence generated within the relationship contexts. However, we assumed in that work that benevolence was purely dispositional, and did not account for a mutualistic form of benevolence as we do in this thesis. Also, we assumed that the ability to reciprocate a goodwill intention from the other partner was also dispositional. Finally, the model of behavior of consumers was too simplistic, where consumers were either benevolent or not benevolent.

The socio-cognitive model of trust (Castelfranchi et al., 2003; Castelfranchi and Falcone, 2010) that we reviewed in Section 3.6, despite being very rich in terms of cognitive construction, is presented in a simplified form in current implementation by the authors' team. Even though, we believe that it would be useful to evaluate the model using populations of agents that evolve with time, situation and relationship – such as the ones generated with the model that we propose in this section – in order to better understand how the model understands the *relationships* that form between trusters and trustees and acts upon this knowledge.

Finally, the model of Adali et al. (2011) that we reviewed in Section 3.6 is the one that could be considered closer to the SOLUM framework. However, as we have mentioned before, this model was not yet properly implemented into a computational approach. Once again, we believe that its evaluation (once it is implemented) using the model of agents' behavior proposed in this section could be valuable.

5.4.2 The Model

The model of agents' behavior described in this subsection is part of the broader agent-based selection scenario described in Section 5.1.1. For simplicity, we consider that there is only one task being negotiated by all trusters, and that all providers accept to negotiate with all consumers. This model starts after the establishment of an agreement between the client and the selected provider, thus excluding the selection process itself. It focus on both type of agents' decision concerning the fulfillment of the agreement: the providers may opt to fulfill it (customers will report outcome F), or to delay its realization; accordingly, the clients may respond to this delay by either retaliating, denouncing the breach (reporting outcome V), or forgiving the contingency (reporting outcome Fd). The behavior of agents at this point is guided by their current benevolence toward the partner, as defined later in this section.

We start the description of our agents' model by defining the main objects of our model. Whenever necessary, we use some formalisms of relational logic to describe concepts that may present some ambiguity. In this regard, we consider that clients (represented by `c`) establish agreements (`a`) with providers (`p`) trusting them to perform a given task (`t`). The complexity and deadline of each task are randomly assigned at setup following a uniform distribution over set $V = \{low, medium, high\}$.

Task Effort. The effort required to successfully perform a given task task is a function of its complexity and deadline. Table 5.5 shows the co-domain of this function given parameters complexity and deadline. $\texttt{effort}(\texttt{t}, \tau)$ denotes that task `t` has effort τ.

Ability of Providers. Following Proposition 8, the ability of agents depends on their individual characteristic and the task itself, and hence is not easily translated into any mathematical distribution representing human populations. We modeled the ability of providers (represented by `ability`) as a random discrete variable taking values in $U = \{very\ low, low, medium,$

Table 5.5: Values for the task required effort given its complexity and deadline.

	Complexity		
Deadline	*low*	*medium*	*high*
low	medium	high	high
medium	low	medium	high
high	low	low	medium

Table 5.6: Probability density function of random discrete variable X.

u	*v. low*	*low*	*medium*	*high*	*v. high*
$P(X = u)$	0.10	0.20	0.40	0.20	0.10

high, *very high*}, with the probability density function (PDF) shown in Table 5.6.

Dispositional Benevolence. Following Proposition 11, agents have a specific disposition to benevolence that is related to their traits of personality. In our research, we could not find a distribution of human dispositional benevolence, as it is tied to the complex concepts of Neuroticism and Agreeableness, which are subject to a variety of development influences (Srivastava et al., 2003). Instead, existing empirical data on benevolence is focused on small homogeneous populations, mostly university students that participate in academic projects. For this reason, we opted to consider that the dispositional benevolence of both clients and providers (represented by `disposition`) is randomly chosen at setup following an uniform distribution over the values in set V.

Satisfaction with the Relationship. Following Proposition 14, the satisfaction of exchange partners increases with the perspective of continuity of the relationship and decreases with the perception of an inequity. We model the perspective of continuity of the relationship, as estimated by any one of the partners, as a function of the trend of interactions between both partners. In turn, the perception of inequities is modeled differently for clients and providers. We consider that a provider defaults when he delays the task at hand; hence, the perception of an inequity by his partner is given by the ratio of the trend of the provider faults (delays) to the trend of the partners' past interactions. On the other hand, a client defaults when he denounces a delay; hence, the perception of inequity by the provider is given by the ratio of the trend of the client's faults (denounces) to the trend of his own faults. The possible values of satisfaction (represented by `satisfaction`) are given

Table 5.7: Satisfaction values given the perspective of continuity and perception of inequity.

Perception of inequity	Perspective of continuity			
	neutral	*low*	*medium*	*high*
neutral	neutral	medium	high	high
low	neutral	medium	medium	high
medium	neutral	low	low	medium
high	neutral	low	low	low

in Table 5.7.

Mutualistic Benevolence. Proposition 15 states that mutualistic benevolence (represented by `mutualistic`) increases with the satisfaction and the exchange compensation and decreases with the number of alternate relationships. We relate the exchange compensation to the effort required to perform the task: smaller efforts bring less risk to consumers. Equations 5.9-5.11 model the mutualistic benevolence of consumers. In this case, we do not consider the existence of alternate relationships because the model assumes that clients are concerned with establishing just one agreement at every simulation round.

$$\begin{aligned} &\texttt{satisfaction}(\texttt{c},\texttt{p},\texttt{neutral}) \wedge \neg\texttt{effort}(\texttt{t},\texttt{low}) \\ &\quad \vee \quad \texttt{satisfaction}(\texttt{c},\texttt{p},\texttt{low}) \Rightarrow \texttt{mutualistic}(\texttt{c},\texttt{p},\texttt{low}) \end{aligned} \tag{5.9}$$

$$\begin{aligned} &\texttt{satisfaction}(\texttt{c},\texttt{p},\texttt{neutral}) \wedge \texttt{effort}(\texttt{t},\texttt{low}) \\ &\quad \vee \quad \texttt{satisfaction}(\texttt{c},\texttt{p},\texttt{medium}) \wedge \neg\texttt{effort}(\texttt{t},\texttt{low}) \\ &\quad \vee \quad \texttt{satisfaction}(\texttt{c},\texttt{p},\texttt{high}) \wedge \texttt{effort}(\texttt{t},\texttt{low}) \\ &\quad \Rightarrow \texttt{mutualistic}(\texttt{c},\texttt{p},\texttt{medium}) \end{aligned} \tag{5.10}$$

$$\begin{aligned} &\texttt{satisfaction}(\texttt{c},\texttt{p},\texttt{medium}) \wedge \texttt{effort}(\texttt{t},\texttt{low}) \\ &\quad \vee \quad \texttt{satisfaction}(\texttt{c},\texttt{p},\texttt{high}) \wedge \neg\texttt{effort}(\texttt{t},\texttt{low}) \\ &\quad \Rightarrow \texttt{mutualistic}(\texttt{c},\texttt{p},\texttt{high}) \end{aligned} \tag{5.11}$$

Equations 5.12-5.14 model the assessment of the providers' mutualistic benevolence. The existence of alternate partners is given by the activity of providers (represented by `activity`), as the slope of the cumulative number of all agreements established by them in the last Δt period of time. Also, providers value more (and then are more benevolent in the presence of) tasks that require more effort, as they may bring higher (monetary/social) compensations.

$$\begin{aligned} &(\texttt{satisfaction}(\texttt{p},\texttt{c},\texttt{neutral}) \vee \texttt{satisfaction}(\texttt{p},\texttt{c},\texttt{low})) \wedge \neg\texttt{effort}(\texttt{t},\texttt{high}) \\ &\quad \vee \quad \texttt{satisfaction}(\texttt{p},\texttt{c},\texttt{medium}) \wedge \texttt{activity}(\texttt{p},\texttt{high}) \Rightarrow \texttt{mutualistic}(\texttt{p},\texttt{c},\texttt{low}) \end{aligned} \tag{5.12}$$

Table 5.8: Total benevolence of agents.

Dispositional	Mutualistic benevolence		
benevolence	*low*	*medium*	*high*
low	very low	low	medium
medium	low	medium	high
high	medium	high	very high

Table 5.9: Ability in agreement.

	Benevolence				
Ability	*v. low*	*low*	*medium*	*high*	*v. high*
very low	very low	very low	low	low	medium
low	very low	low	low	medium	medium
medium	low	low	medium	medium	high
high	low	medium	medium	high	very high
very high	medium	medium	high	high	very high

$$\begin{aligned}&(\texttt{satisfaction}(\texttt{p},\texttt{c},\texttt{neutral}) \vee \texttt{satisfaction}(\texttt{p},\texttt{c},\texttt{low}))\\ &\quad\wedge\ \texttt{effort}(\texttt{t},\texttt{high}) \wedge (\texttt{activity}(\texttt{p},\texttt{medium})\\ &\quad\vee\ \texttt{activity}(\texttt{p},\texttt{high})) \vee \texttt{satisfaction}(\texttt{p},\texttt{c},\texttt{medium}) \wedge \neg\texttt{activity}(\texttt{p},\texttt{high})\\ &\quad\vee\ \texttt{satisfaction}(\texttt{p},\texttt{c},\texttt{high}) \wedge \neg\texttt{effort}(\texttt{t},\texttt{high}) \wedge (\texttt{activity}(\texttt{p},\texttt{medium})\\ &\quad\vee\ \texttt{activity}(\texttt{p},\texttt{high})) \Rightarrow \texttt{mutualistic}(\texttt{p},\texttt{c},\texttt{medium})\end{aligned} \tag{5.13}$$

$$\begin{aligned}&\texttt{satisfaction}(\texttt{p},\texttt{c},\texttt{high}) \wedge \neg\texttt{effort}(\texttt{t},\texttt{high}) \wedge (\texttt{activity}(\texttt{p},\texttt{neutral})\\ &\quad\vee\ \texttt{activity}(\texttt{p},\texttt{low})) \vee \texttt{satisfaction}(\texttt{p},\texttt{c},\texttt{high}) \wedge \texttt{effort}(\texttt{t},\texttt{high})\\ &\quad\vee\ \neg\texttt{satisfaction}(\texttt{p},\texttt{c},\texttt{high}) \wedge \texttt{effort}(\texttt{t},\texttt{high}) \wedge (\texttt{activity}(\texttt{p},\texttt{neutral})\\ &\quad\vee\ \texttt{activity}(\texttt{p},\texttt{low})) \Rightarrow \texttt{mutualistic}(\texttt{p},\texttt{c},\texttt{high})\end{aligned} \tag{5.14}$$

Total Benevolence. The possible values of total benevolence of both clients and providers are given in Table 5.8.

Ability in Agreement. We consider that providers can present a little more ability in specific agreements if they are highly benevolent toward their partners – e.g., by outsourcing some of the effort required to perform the task – and less ability if their benevolence is low, by putting less effort to the task than their real ability. Table 5.9 presents the ability of providers in an agreement given their innate ability and their total benevolence toward the exchange partner. This 'modulated' ability is represented by term `abilityIn`.

Delay. Providers delay their agreements (represented by $\texttt{delay}(\texttt{p},\texttt{a})$) when their ability in the agreement is not enough to handle the required effort

Table 5.10: Probabilities of denouncing after a breach of agreement.

	Mutualistic benevolence		
Dispositional benevolence	*low*	*medium*	*high*
low	0.9	0.6	0.2
medium	0.7	0.4	0.2
high	0.5	0.2	0.1

of the task at hands (cf. equations 5.15 and 5.16). In the equations, `largOrEqual` is a predicate symbol that yields true when the value of the first argument is equal or bigger than the value of the second argument.

$$\forall \tau_1,\tau_2 \quad \texttt{largOrEqual}(\texttt{abilityIn}(\texttt{p},\texttt{a},\tau_1),\texttt{effort}(\texttt{t},\tau_2))\neg\texttt{delay}(\texttt{p},\texttt{a}) \tag{5.15}$$

$$\forall \tau_1,\tau_2 \quad \neg\texttt{largOrEqual}(\texttt{abilityIn}(\texttt{p},\texttt{a},\tau_1),\texttt{effort}(\texttt{t},\tau_2)) \Rightarrow \texttt{delay}(\texttt{p},\texttt{a}) \tag{5.16}$$

Denounce. Clients have a probability to denounce after suffering a breach of an agreement that is given by their dispositional and mutualistic benevolence, as shown in Table 5.10.

Having described the behavioral model of agents, we are in conditions of describing the first set of experiments we performed to evaluate the *Social Tuner* component, what we do next.

5.5 Evaluation of *Integrity Tuner*

In this section, we evaluate Hypothesis 3, which says that the extraction of integrity-based information from the set of evidence on the trustee under evaluation improves the reliability of the estimation of this trustee's trustworthiness. This also partially addresses Research Question 1, that considers the need to distinguish between the different trustworthiness dimensions in order to get more reliable trustworthiness estimations.

For this, we ran different experiments in our agent-based simulated environment where, at every round of the experiments, different types of trusters chose the best partners to perform a task from a set of trustees with different characteristics. For simplicity, we considered that there was only one task being negotiated by all trusters, although its requirements in terms of complexity and deadline changed with round and truster; also, all trustees accepted to negotiate with all trusters. Moreover, we used the behavioral model of agents described in Section 5.4. This model ran just after the establishment of an agreement between any given truster and the selected trustee, thus excluding the selection process itself. It focus on both types of agents' decision concerning the fulfillment of the established agreement:

the trustees may opt to fulfill the agreement (trusters will report outcome F), or to delay its realization; accordingly, the trusters may respond to a delay by either retaliating, denouncing the breach (reporting outcome V), or forgiving the contingency (reporting outcome Fd).

5.5.1 First Set of Experiments

In this set of experiments, we wanted to test the following hypothesis:

Hypothesis 9 *In the presence of populations of trusters and trustees that evolve their behavior based on the relationships they are able to develop with others, trusters that are able to extract the integrity of the trustees from the available evidence using* Integrity Tuner *will perform better than those that do not have this ability.*

With the intent of testing this hypothesis, we compared the use of our situation-less trustworthiness estimator *Sinalpha* (model `S`) with a trust model consisting of the joint use of *Sinalpha* and *Integrity Tuner* (model `SI`). The estimated trustworthiness of agents of type `S` was then calculated based only on the situation-less form of ability, such that $tw_{x,y} = Sinalpha(E_{*,y})$. In turn, the estimated trustworthiness of `SI` agents was calculated using both the estimated situation-less ability of the trustee and the estimated value of his integrity, such that $tw_{x,y} = 1/2 Sinalpha(E_{*,y}) + 1/2 IntegrityTuner(E_{*,y})$. We used a third population of trusters, which were not able to estimate the trustworthiness of trustees (model `NT`). Hence, model `NT` acted as a baseline for evaluation purposes.

Selection Criteria

`NT` agents selected their partners based only on the utility of the providers' proposals, such that $D_x^{up} = \arg\max_{y_i \in \mathcal{Y}}(up_{y_i})$. `S` agents used a dual selection criteria – the trustees' estimated trustworthiness and the utility of their proposals, such that $D_x^{tr-ut} = \arg\max_{y_i \in Y}(1/2 tw_{x,y_i} + 1/2 up_{y_i})$. Finally, `SI` agents used the same dual selection criteria, but added the additional procedure illustrated in Algorithm 1.

This way, clients of type `SI` ordered all received proposals in descendant order, using criteria $(1/2 tw_{x,y_i} + 1/2 up_{y_i})$. Then, they further tested if the first ranked proposal was sent by a provider of integrity. If it was the case, this was the selected proposal (line 15). Otherwise, the proposal was removed from the list (lines 12-14), and a similar process was applied to the second best proposal, and then to the third, and so one, until a proposal

Algorithm 1 Additional selection procedure for agents of type `SI`.

```
1:  function ADD_SEL_PROC_SI (P) returns p
2:      P: the set of all proposals, ordered by trustworthiness and utility
3:      ml: minimum limit for considering integrity
4:      cst: consistency threshold
5:
6:      firstP ← P[0]
7:      while |P| >= 1
8:          p ← P[0]
9:          np ← no. of past agreements of p's proponent
10:         cs ← estimated consistency of p's proponent
11:         int ← estimated integrity of p's proponent
12:         if np ≥ 1 and np < ml and (cs = 0 or int = 0) then remove(p)
13:         else if np ≥ ml and int = 0 then remove(p)
14:         else if np ≥ ml and cs < cst then remove(p)
15:         else return p
16:     return firstP
```

was selected or there were no more proposals to analyze (line 7). At the end of the process, if none of the analyzed proposals were from providers of integrity, the first proposal of the original ordered set was selected (lines 6 and 16).

In the process just described, we used rule-based heuristics to decide if any given provider was an agent of integrity. These rules resulted from an extensive analysis of the results of multiple experiments with *Integrity Tuner*, and also from the insights derived from the literature revision on integrity. For instance, Proposition 18 refers that the integrity of an individual may be perceived by a partner earlier in the relationship. In the same way, Dunn and Schweitzer (2005) refer that when people judge trust in acquaintances, they must probably use a heuristic information-processing strategy. Taking these insights into consideration, we defined the *minimum limit for integrity* (ml, line 3), which indicates the number of individual items of evidence that are necessary to reason about the agents' integrity. In these experiments, we used our common sense to consider that $ml = |\mathcal{O} + 2|$. Then, if the evidence size on the provider under assessment was below ml, we could do no better than realize that if the estimated consistency of the actions of this provider (as estimated by coefficient ρ_{cs} of *Integrity Tuner*, line 10) was zero, this provider generated different outcomes in all of his past agreements, indicating that probably he was not an agent of integrity; hence, his proposal was removed from selection (line 12). In the same way, if his estimated integrity

Table 5.11: Configuration parameters (evaluation of *Integrity Tuner*).

terms of the cfp	randomly assigned at each round
$\mathcal{V}_6$	*low*, *medium*, *high*
$\mathcal{V}_7$	*low*, *medium*, *high*
$\mathcal{O}$	F, Fd, V
terms of proposals	same as in cfp
selection criteria	depends on truster type
up_y	uniform distribution over $[0.5, 1.0]$
clients' dispositional benevolence	*medium*
providers' ability	cf. Table 5.6
providers' disp. benevolence	un. dist. over $\{low, medium, high\}$
#clients, #providers	24, 20
#rounds, #exp. repetitions	100, 30
Sinalpha ω, λ_F, $\lambda_F d$, λ_V	$\pi/12$, 1.0, -0.5, -2.0
Sinalpha $lf(F)$, $lf(Fd)$, $lf(V)$	0.0, 0.5, 1.0
`B` and `J` parameters	cf. Section 5.3.1
$vlr(F)$, $vlr(Fd)$, $vlr(V)$	1.0, 0.5, -0.5
ml, cst	5, 0.5

(as calculated by *Integrity Tuner*, line 11) was zero, this indicates that the provider failed all past agreements, and therefore his proposal was removed (line 12). Of course, this last rule was also valid if the evidence set on the provider was larger than ml (line 13).

Finally, we verified from the traces of the experiments (using *Sinalpha* and two other trust-based evidence aggregators) that providers that already established several agreements in the past tended to maintain moderate to high values of trustworthiness, even if they occasionally delayed or violated these agreements. Hence, we established a *consistency threshold* (cst, line 4) and compared the consistency of providers against this threshold; providers showing consistency values lower than cst saw their proposals removed from selection (line 14), allowing for the search of putative more consistent providers that, for some valid reason, could have established less agreements in the past.

Configuration Parameters

In this set of experiments, we ran 8 clients of type `NT`, 8 clients of type `S`, and 8 clients of type `SI`. In order to attenuate the effect of benevolence in these particular experiments, all client agents were set with a medium value of dispositional benevolence, at setup. Table 5.11 presents the configuration parameters of these experiments.

Table 5.12: Results of the first set of experiments with *Integrity Tuner*.

	F		V		D		U	
	M	SD	M	SD	M	SD	M	SD
NT	0.552	0.069	0.248	0.040	0.842	0.010	0.979	0.003
S	0.825	0.104	0.094	0.058	0.299	0.073	0.829	0.025
SI	0.860	0.101	0.073	0.056	0.300	0.106	0.829	0.041

Results

Table 5.12 summarizes the results obtained in this set of experiments. It shows the mean values (M) and standard deviation (SD) of variables F (average percentage of outcomes of type F), V (average percentage of outcomes of type V), D (average percentage of different providers selected by all clients at one round), and U (average utility of the selected proposals, at one round).

We verified that agents of type S had more 49.40% of agreements with outcome F than agents of type NT, and that agents that additionally used *Integrity Tuner* (SI agents) got 55.80% more of these agreements than agents of type NT. In the same way, the number of violated agreements when compared to the results of NT agents reduced 62.01% for S agents and 70.74% for agents of type SI.

In terms of the mean utility of the selected proposals, we verified that the impact of the addition of the integrity-based functionalities was neutral, as there were no statistically significant differences in the results obtained by S (M: 0.829, SD: 0.025) and SI agents (M: 0.829, 0.041), $t(29) = 0.090, p > 0.02$. The same applies to the mean number of different providers selected at each round, which was similar for S (M: 0.299, SD: 0.073) and SI agents (M: 0.300, 0.106), $t(29) = -0.119, p > 0.02$.

Discussion

Based on the obtained results using the *Sinalpha* trust-based evidence aggregator, we are able to confirm the truthfulness of Hypothesis 9, in the conditions of the experiments. In fact, trusters that were able to consider the consistency and the integrity of the candidate partners made wiser decisions, which reflected in an increased number of successful agreements, with no loss of utility.

It is worth to note that these results were obtained using populations of agents that were not characterized by their inherent integrity, which makes us believe that in fact the use of *Integrity Tuner* improves the reliability of the estimation of the agents' trustworthiness (cf. Hypothesis 3).

Table 5.13: Results of the second set of experiments with *Integrity Tuner*.

	F		V		D		U	
	M	SD	M	SD	M	SD	M	SD
S	0.800	0.139	0.105	0.077	0.465	0.096	0.829	0.035
SI	0.835	0.146	0.091	0.081	0.447	0.137	0.822	0.048
B	0.779	0.093	0.124	0.055	0.664	0.065	0.900	0.020
BI	0.820	0.127	0.102	0.075	0.514	0.151	0.848	0.054
J	0.801	0.096	0.105	0.056	0.599	0.080	0.876	0.027
JI	0.853	0.130	0.083	0.079	0.478	0.158	0.833	0.054

5.5.2 Second Set of Experiments

In this second set of experiments, we wanted to test the following hypothesis:

Hypothesis 10 *The benefits of the* Integrity-Tuner *component can be shown when applied to different types of trustworthiness estimators.*

In order to test this hypothesis, we applied *Integrity Tuner* to the computational trust models B and J, already described when evaluating the *Contextual Fitness* component, in Section 5.3.1. Hence, we ran six different types of clients simultaneously, each with four agents: S, SI, B, BI (combining B with *Integrity Tuner*), J, and JI (combining J with *Integrity Tuner*). All other configuration parameters used in the first set of experiments were maintained.

Results

Table 5.13 summarizes the results obtained in this set of experiments. As can be observed in this table, all trust-based aggregation engines profited from the inclusion of *Integrity Tuner* in terms of the number of agreements with outcome F. In fact, the addition of the integrity-based functionalities in the decision making of agents of type S, B, and J improved this number in 4.33%, 5.26%, and 6.49%, respectively. In the same way, the same type of agents saw the number of violated agreements reduced in 13.38%, 18.01%, and 21.02%, respectively, when compared to their integrity-less counterparts.

However, these good results came at the price of a slight reduction in the mean utility of the selected proposals for agents of type BI and JI (variable U), as they tended to stick with the same, more integer providers (thus being less exploratory, as indicates variable D), which might not be the ones proposing the most useful proposals.

Discussion

The results of these experiments showed that the addition of *Integrity Tuner* to different computational trust approaches proposed in literature increased the reliability of their trustworthiness estimations. Hence, Hypothesis 10 is accepted in the conditions of the experiments.

5.5.3 Third Set of Experiments

In the third set of experiments, we wanted to further test hypothesis 9 in populations that would show some kind of integrity-based characteristics. Hence, we slightly changed the population of provider agents, by ascribing them a disposition to integrity value, randomly assigned at setup over values $\{low, medium, high\}$. This disposition affected the way these agents made their proposals: providers of low integrity offered proposals with high utility to clients ($\mu_p \in [0.6, 1.0]$), eagerly seeking to be selected, despite their abilitt; providers of medium integrity offered proposals with medium utility to clients ($\mu_p \in [0.4, 0.8]$); and providers of high integrity offered proposals with lower utility to clients ($\mu_p \in [0.2, 0.6]$). Although it is realistic to think that providers with high integrity may offer high utility to the clients and the other way around, we wanted with this setting to guarantee that the proposals of providers of low integrity tended to be ranked in the first positions, following the chosen selection criteria. The remaining conditions settle in the second set of experiments were maintained.

Results

Table 5.14 summarizes the results obtained in this set of experiments. Once again, all trust-based aggregation engines profited from the inclusion of *Integrity Tuner* in terms of the number of agreements with outcome F. In fact, the addition of the integrity-based functionalities in the decision making of agents of type `S`, `B`, and `J` increased the of outcomes of this type in 4.95%, 10.59%, and 8.64%, respectively. In the same way, it significantly reduced the number of violated agreements in 29.11%, 45.86%, and 44.16%, respectively.

It is interesting to note that all six types of clients increased the number of successful agreements and decreased the number of violated agreements. This happened because the integrity-less clients profitted from the good choices of their integrity-aware counterparts.

Table 5.14: Results of the third set of experiments with *Integrity Tuner*.

	F		V		D		U	
	M	SD	M	SD	M	SD	M	SD
S	0.855	0.118	0.079	0.067	0.359	0.081	0.736	0.101
SI	0.897	0.112	0.056	0.065	0.322	0.081	0.686	0.146
B	0.803	0.127	0.113	0.077	0.493	0.114	0.845	0.040
BI	0.888	0.116	0.061	0.065	0.354	0.106	0.697	0.146
J	0.829	0.113	0.091	0.065	0.446	0.104	0.823	0.043
JI	0.901	0.098	0.051	0.054	0.331	0.091	0.688	0.145

5.6 Evaluation of *Social Tuner*

In order to evaluate the contribution of our approach concerning Research Question 3, more specifically, the possibility of computational trust models to infer from the available evidence on a given trustee the relationships existing between this trustee and his evaluators, we ran a set of experiments with *Social Tuner*. In particular, we wanted to evaluate if the use of *Social Tuner* increased the performance of the trustworthiness estimator in the presence of populations of truster and trustees with different abilities and able of developing benevolent relationships between them (Hypothesis 1).

Contrary to what happened when evaluating the *Integrity Tuner* component, the dispositional benevolence of clients in this set of experiments was randomly chosen over set $\{low, medium, high\}$. Table 5.2 shows the configuration parameters that are common to all sets of experiments with *Social Tuner*.

5.6.1 First Set of Experiments

In this set of experiments, we wanted to test the following hypothesis:

Hypothesis 11 *In the presence of populations of trusters and trustees that evolve their behavior based on the benevolent relationships they are able to develop with each others, trusters that are able to extract the benevolence of the trustees toward the trusters from the available evidence using* Social Tuner *will perform better than those that do not have this ability.*

With the intent of testing this hypothesis, we compared the use of our situation-less trustworthiness estimator *Sinalpha* (model S) with a trust model consisting of the joint use of *Sinalpha* and *Social Tuner* (model SB). We used a third population of trusters, which were not able to estimate the

Table 5.15: Configuration parameters (evaluation of *Social Tuner*).

terms of the cfp	randomly assigned at each round
$\mathcal{V}_6$	*low*, *medium*, *high*
$\mathcal{V}_7$	*low*, *medium*, *high*
$\mathcal{O}$	F, Fd, V
terms of proposals	same as in cfp
selection criteria	depends on truster type
up_y	uniform distribution over $[0.5, 1.0]$
#clients, #providers	24, 20
# rounds	varies with experiment
# experiment repetitions	30
clients' dispositional benevolence	un. dist. over $\{low, medium, high\}$
providers' ability	cf. Table 5.6
providers' disp. benevolence	un. dist. over $\{low, medium, high\}$
$vl(F)$, $vl(Fd)$, $vl(V)$	1.0, 0.5, 0.0
Sinalpha ω, λ_F, $\lambda_F d$ and λ_V	$\pi/12$, 1.0, -0.5 and -2.0
Sinalpha $lf(F)$, $lf(Fd)$ and $lf(V)$	0, 0.5 and 1

trustworthiness of trustees (model `NT`). Hence, model `NT` acted as a baseline for evaluation purposes.

Trusters of type `S` estimated the trustworthiness of trustees ($tw_{x,y}$) in the same way as when evaluating *Integrity Tuner*. However, trusters of type `SB` estimated the trustworthiness of trustees using the trustworthiness evaluation function shown in Algorithm 2. This function takes into consideration the truster's perception of the ability and benevolence of the trustee and weights both dimensions according to the relationship existing between truster and trustee.

In the algorithm above, we measured the number of interactions between x and y, $N_{x,y}$ (line 8), and defined a minimum number of interactions between truster x and trustee y, $N_{ben_{close}}$, after which the partners were considered to be engaged in a close relationship (lines 3 and 9). Also, we considered a weight of benevolence, ω_{ben}, to be used when combining the estimated value of the trustee's ability as returned by *Sinalpha* (line 6) with the estimated value of its benevolence as returned by *Social Tuner* (line 7). This weight was set to zero when there was just one or zero interactions between both partners (line 11), and then progressively increased with the growing number of interactions between the partners, until it reached the maximum value of one when the partners were considered to be in a close relationship (line 10). Finally, the estimated value of the trustee's trustworthiness ($tw_{x,y}$) was computed using the weighted mean of $ab_{a,y}$ and $ben_{x,y}$

Algorithm 2 Computation of $tw_{x,y}$ for agents of type `SB`.

1: **function** TW_SB ($E_{*,y}, N_{ben_{close}}$) returns $tw_{x,y}$
2: $E_{*,y}$: the set of all evidence about trustee y
3: $N_{ben_{close}}$: minimum (x, y) interactions for closeness
4:
5: $E_{x,y} \leftarrow \{e_i \in E_{*,y} \ : \ v_1^{e_i} = x\}$
6: $ab_{x,y} \leftarrow$ *Sinalpha* $(E_{*,y})$
7: $ben_{x,y} \leftarrow$ *Social Tuner* $(E_{x,y})$
8: $N_{x,y} \leftarrow |E_{x,y}|$
9: **if** $N_{x,y} > N_{ben_{close}}$ **then** $N_{x,y} = N_{ben_{close}}$
10: **if** $N_{x,y} > 1$ **then** $\omega_{ben} = N_{x,y}/N_{ben_{close}}$
11: **else** $\omega_{ben} = 0$
12: $tw_{x,y} = (1 - \omega_{ben}) \cdot ab_{x,y} + \omega_{ben} \cdot ben_{x,y}$
13: **return** $tw_{x,y}$

with weights $(1 - \omega_{ben})$ and ω_{ben} (line 12), respectively.

By considering a progressive use of the estimated benevolence of the trustee, we took into consideration the perception of individual trustworthiness dimensions and the relative importance of each one of them, as postulated in propositions 18 and 19.

Selection Criteria

Trusters of type `NT` selected the best providers based on the utility of their proposals, making $D_x^{up} = \arg\max_{y_i \in Y}(up_{y_i})$. Trusters of type `S` used a dual selection criteria – the trustees' estimated trustworthiness and the utility of their proposals ($D_x^{tr\_ut}$) –, using the following formula: $D_x^{tr\_ut} = \arg\max_{y_i \in Y}(1/2tw_{x,y_i} + 1/2up_{y_i})$. Trusters of type `SB` used the same dual selection criteria, but added the additional procedure described as follows: just before ordering the proposals by trustworthiness and utility, the truster removed from the set of all considered proposals these proposals owned by trustees that did not reach a given benevolence threshold. We considered and tested three different thresholds: the mean of all the values of benevolence shown by the trustees (`mean`); the maximum of all of these values (`max`); and the average of the mean and the maximum benevolence values (`mean-max`).

In this set of experiments, we ran 8 clients of type `NT`, 8 clients of type `S`, and 8 clients of type `SB`. In order to better evaluate the effect of using the *Social Tuner* component in different conditions regarding the number of interactions between trusters and trustees, we further ran the experiments

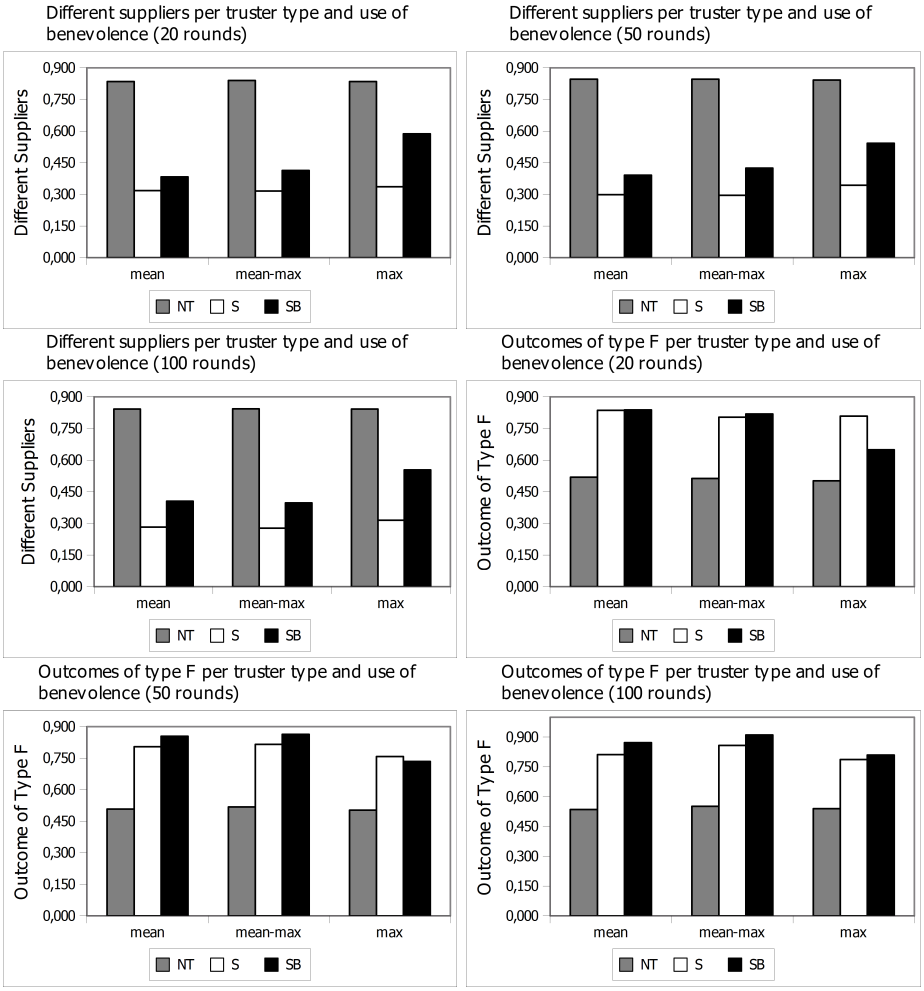

Figure 5.9: Results (different suppliers and outcomes of type F) from the evaluation of *Social Tuner*, per truster type (first set of experiments).

with 20 rounds, 50 rounds, and 80 rounds. We set $N_{ben_{close}} = 15$.

Results

The results of all experiments for the three considered thresholds are shown in figures 5.9 and 5.10. Starting with the exploration tendency of each truster type (Figure 5.9, *top* and *middle-left*), we observed that the use of *Social Tuner* increased the number of different selected providers for all configurations of the number of rounds and benevolence thresholds. From all trusters of type `SB`, those of type `max` were notably the ones that interacted with more partners.

The effect of `SB` trusters exploring more than `S` trusters on the number of

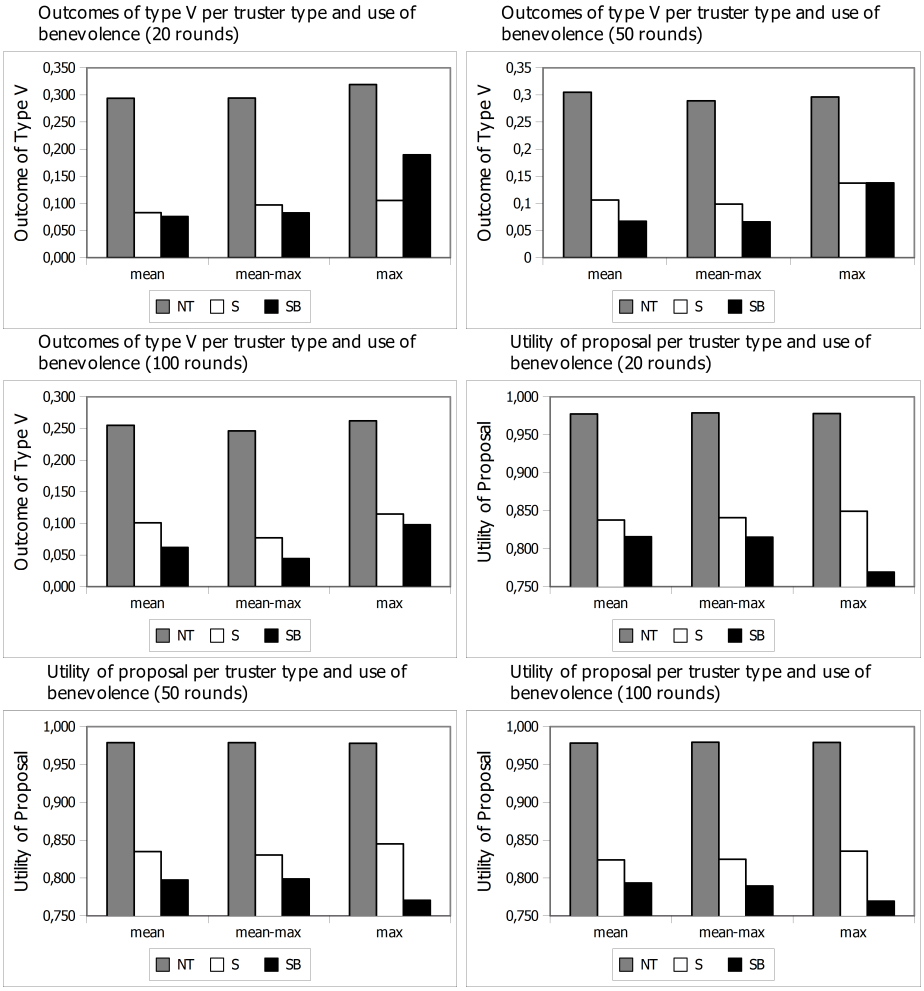

Figure 5.10: Results (outcomes of type V and utility of proposals) from the evaluation of *Social Tuner*, per truster type (first set of experiments).

outcomes of type F (Figure 5.9, *middle-right* and *bottom*) depended on the used benevolence threshold and number of rounds considered. For instance, with only 20 rounds, when the number of interactions between any two partners was not large, only the trusters of type `mean-max` got more outcomes of type F than trusters of type `S`, although the difference was not statistically significant when using Bonferroni adjustment: `S` ($M = 0.803, SD = 0.080$), $\texttt{SB}_{mean-max}$ ($M = 0.818, SD = 0.075$), $t(29) = -1.73$, $p = 0.05$. In turn, and also considering 20 rounds, the trusters of type `max` got significantly less outcomes of type F than trusters of type `S`, which indicates that the exclusion of a great number of trustees since the first rounds, where most of these trustees had not yet have the chance to show their benevolence, is

a poor benevolence-based selection approach. The benefits of using the `max` threshold were only observed when considering 100 rounds.

However, the use of both `mean` and `mean-max` benevolence thresholds allowed for an increase of outcomes of type F when compared to trusters of type `S`, for 50 and 100 rounds. For instance, when comparing trusters of types `S` and $\texttt{SB}_{mean-max}$ considering 100 rounds, the latter had 9.42% more outcomes of type F than the former when compared to baseline trusters `NT`, `S` ($M = 0.858, SD = 0.093$), $\texttt{SB}_{mean-max}$ ($M = 0.910, SD = 0.053$), $t(29) = -5.06$, $p < 0.017$, indicating that *Social Tuner* was being effective in capturing the benevolence existing between any truster-trustee pair. It is worth noting that the evolving behavior of agents as modeled in Section 5.4 depended on other factors beside benevolence, such as the value of the agreement under consideration and the trustees' ability, so we considered that the results obtained with trusters of type $\texttt{SB}_{mean-max}$ in the variable being analyzed were very promising.

The results regarding the number of outcomes of type V per truster (Figure 5.10, *top* and *middle-left*) confirmed the good results of truster types `mean` and `mean-max`. For instance, when comparing trusters of types `S` and $\texttt{SB}_{mean-max}$ considering 100 rounds, the latter got 13.28% less outcomes of type V than the former when compared to baseline trusters `NT`, with `S` ($M = 0.077, SD = 0.057$), $\texttt{SB}_{mean-max}$ ($M = 0.044, SD = 0.028$), $t(29) = 4.98$, $p < 0.017$. From all performed experiments, the `mean-max` benevolence threshold seemed to be the one that allowed for a better balance when combining the estimated values of ability and benevolence, allowing for better results in terms of outcomes of type F and V, in all used configuration of rounds.

Finally, when comparing the average utility of proposal per truster type (Figure 5.10, *middle-right* and *bottom*), we observed that trusters of type `SB` got slightly less utility than trusters of type `S`. For instance, when comparing trusters of types `S` and $\texttt{SB}_{mean-max}$ considering 100 rounds, the latter got 3.61% less average utility per proposal than the former, when compared to baseline trusters `NT`: `S` ($M = 0.825, SD = 0.021$), $\texttt{SB}_{mean-max}$ ($M = 0.789, SD = 0.018$), $t(29) = 9.16$, $p < 0.017$. This results was as expected, as trusters of type $\texttt{SB}_{mean-max}$ were more effective than trusters of type `S` in selecting the most trustworthy trustees, given their relationship with trusters, and those trustees might not be the ones that presented the highest utility, at every round. However, when compared to the gains achieved with the use of the *Social Tuner* in terms of outcomes of types F and V, we believe that the loss of about 3% in utility is irrelevant.

Discussion

We verified that the selection of the proposals yielded best results in terms of F and V outcomes when the benevolence of the proponents were above the `mean-max` benevolence threshold. Overall, in the conditions of these experiments, we were able to confirm the truthfulness of Hypothesis 11.

5.6.2 Second Set of Experiments

In this set of experiments, we wanted to test the following hypothesis:

Hypothesis 12 *The benefits of the* Social-Tuner *component can be shown when applied to different types of trustworthiness estimators.*

In order to test this hypothesis, we ran another set of experiments, where *Social Tuner* was applied to the computational trust models `B` and `J`, already described when evaluating the *Contextual Fitness* and *Integrity Tuner* components. Hence, we ran six different types of trusters simultaneously, each with four agents: `S`, `SB` (combining `S` with *Social Tuner*), `B`, `BB` (combining `B` with *Social Tuner*), `J`, and `JB` (combining `J` with *Social Tuner*). All trusters used the `mean-max` benevolence threshold.

Results

The results of these experiments in terms of outcomes of type F and V are shown in Figure 5.11. Table 5.16 systematizes the results in terms of outcomes of type F, discriminating between mean values (M) and standard deviation (SD). These results confirmed that the addition of *Social Tuner* to the different simple trustworthiness estimators increased the number of outcomes of type F and decreased the number of outcomes of type V for all of these trustworthiness estimators, for all number of rounds considered.

Discussion

In this set of experiments, we went beyond traditional evaluation of computational trust models and developed a model of agents' behavior where both trusters and trustees evolve their behaviors based on personality traits, mutualistic interests and the stage of the different relationships existing between the agents. Using this model, we evaluated the potential benefits of using the *Social Tuner* component over three different simple trustworthiness estimators. The resulting benevolence-enhanced models aggregated the values of the estimated ability (as calculated by the trustworthiness estimators) with the estimated benevolence (as calculated by *Social Tuner*) in

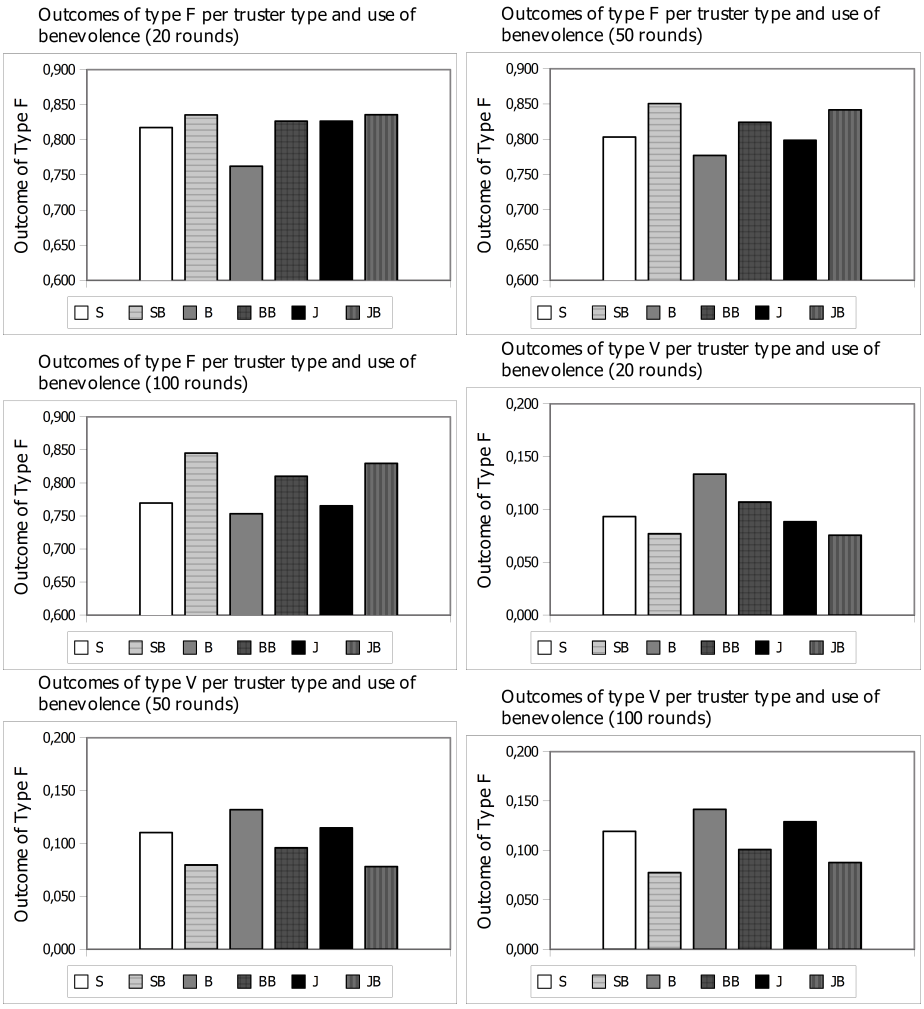

Figure 5.11: Results for *Social Tuner* using different trustworthiness estimators.

Table 5.16: Results of the second set of experiments with *Social Tuner* (variable `F`).

	20 rounds		50 rounds		100 rounds	
	M	SD	M	SD	M	SD
S	0.817	0.081	0.803	0.113	0.770	0.105
SB	0.835	0.067	0.850	0.095	0.845	0.091
B	0.762	0.090	0.777	0.074	0.753	0.069
BB	0.798	0.084	0.824	0.091	0.810	0.070
J	0.827	0.084	0.799	0.075	0.765	0.067
JB	0.836	0.071	0.842	0.081	0.830	0.081

a dynamic way, where the weight of benevolence grew with the increasing number of interactions between any truster-trustee pair. Moreover, we presented an algorithm for the selection of partners where the proposals made by trustees presenting an estimated value of benevolence lower than a given benevolence threshold were not considered.

When we adopted the `mean-max` threshold, we verified that all computational trust models that we tested that added the *Social Tuner* component allowed for significant more reliable trustworthiness estimations than their counterparts without the *Social Tuner* Component. Overall, in the conditions of these experiments, we were able to confirm the truthfulness of Hypothesis 12.

5.6.3 Third Set of Experiments

In this set of experiments, we wanted to test the following hypothesis:

Hypothesis 13 *In the presence of populations of trusters and trustees of homogeneous benevolence, trusters that use the* Social Tuner *component will perform no worse than those that do not use this component.*

In order to test this hypothesis, we made important changes to the behavioral model of agents described in Section 5.4. First, we set the dispositional benevolence of both trusters and trustees to a fixed value of *Medium*. Second, the ability in agreement, which determines whether the trustees fulfilled or delayed their agreements given the effort required to perform the agreement, was no longer dependent on the benevolence of these trustees toward the exchange partners, and was given solely by the trustees' ability. Hence, the resulting agents were not driven by benevolence.

We ran this set of experiments with six different types of trusters running simultaneously, each with four agents: `S`, `SB`, `B`, `BB`, `J`, and `JB`. All trusters used the `mean-max` benevolence threshold, and all experiments had 100 rounds.

Results

The results of these experiments in terms of outcomes of type F and V, considering 100 rounds, are shown in Table 5.17. From the results, we observed that no one of the three chosen trustworthiness estimators (i.e., `S`, `B`, and `J`) performed poorly when combined with the *Social Tuner* component. In fact, all of them performed a little better in terms of outcome F, although this increase was only statistically significant (using Bonferroni

Table 5.17: Results of the third set of experiments with *Social Tuner* (variable F).

	F		V	
	M	SD	M	SD
S	0.941	0.055	0.028	0.028
SB	0.947	0.061	0.025	0.030
B	0.910	0.058	0.048	0.034
BB	0.933	0.059	0.031	0.028
J	0.927	0.056	0.036	0.028
JB	0.938	0.054	0.027	0.026

adjustments) with model B ($t(29) = -4.51, p < 0.003$). The same happened with outcome of type V, where the decrease observed with model B was statistically significant ($t(29) = 5.67, p < 0.003$).

Overall, in the conditions of these experiments, we were able to confirm the truthfulness of Hypothesis 13.

5.7 Evaluation of the $Tw_{x,y}$ Approaches

In this section, we present the experimental evaluation of the two simple approaches instantiating function $Tw_{x,y}$, described in Section 4.8. In order to evaluate these approaches, we used the same testbed that was used to evaluate *Social Tuner*, in the previous section, with little adjustments. First, we added to the characterization of provider agents the integrity-based disposition mentioned when we evaluated the *Integrity Tuner* component, in the third set of experiments of Section 5.5.

Second, we ran six different populations of agents simultaneously, each one representing one or more of the trustworthiness dimensions that we have been evaluating individually throughout this chapter. Hence, we had the following client populations, with four agents each: NT, our baseline population; S, using the trust-based evidence aggregator *Sinalpha*; SC, representing those agents that used *Sinalpha* with *Contextual Fitness*, as evaluated individually in Section 5.3; SI, representing those agents that used *Sinalpha* with *Integrity Tuner*, as evaluated individually in Section 5.5; SB, representing those agents that used *Sinalpha* with *Social Tuner*, as evaluated individually in Section 5.6; and finally All, representing those agents that combined all these situation-based, integrity-based, and benevolence-based features. Each experiment ran in 80 rounds.

Table 5.18: Results of the first set of experiments with $Tw_{x,y}$.

	F		V	
	M	SD	M	SD
NT	0.489	0.097	0.306	0.073
S	0.785	0.158	0.115	0.091
SC	0.795	0.153	0.116	0.097
SB	0.846	0.106	0.075	0.060
SI	0.838	0.136	0.084	0.082
All	0.852	0.119	0.063	0.073

5.7.1 First Set of Experiments

In these experiments, we wanted the evaluate the first alternative to implementing function $Tw_{x,y}$, as described in Section 4.8.1.

Results

The results of these experiments in terms of outcomes of type F and V, considering 80 rounds, are shown in Table 5.18. As was expected, all client types that used trust outperformed population NT. When comparing the resulting outcomes of type F with the baseline population, agents of type S got 60.42% more of these outcomes, SC agents were 62.40% better, SI agents were 71.19% better, SB agents were 72.96% better, and All agents were 74.18% better. However, the difference between agents of types SB and All were not statistically significant.

The results concerning the total number of violated contracts were in line with these results for variable F. Relatively to NT, S reduced the violations in 62.31%, SC reduced the violations in 62.09%, and SI, SB and All had less 72.44%, 75.60% and 79.30% violated agreements, respectively.

Discussion

Although preliminary, these results confirmed our intuition (expressed in Research Question 1) that reasoning about ability, integrity, and benevolence as individual dimensions, and combining them taking into consideration the development of the relationship with the trustees, allowed the trusters to make more informed decisions about the trustees' trustworthiness.

Table 5.19: Results of the second set of experiments with $Tw_{x,y}$.

	F		V	
	M	SD	M	SD
NT	0.517	0.097	0.305	0.083
S	0.869	0.104	0.064	0.052
SC	0.876	0.114	0.063	0.062
SB	0.896	0.081	0.048	0.051
SI	0.909	0.096	0.048	0.055
All	0.921	0.090	0.038	0.040

5.7.2 Second Set of Experiments

In these experiments, we wanted the evaluate the second, heuristic-based alternative for implementing function $Tw_{x,y}$, as described in Section 4.8.2.

Results

The results of these experiments in terms of outcomes of type F and V, considering 80 rounds, are shown in Table 5.19. Once again, we verified that all models that added situation-based and/or social-based features to *Sinalpha* improved the reliability of the trustworthiness estimation of this aggregator, in terms of generated outcomes of type F, and were at least as good as *Sinalpha* in terms of outcomes of type V. For instance, when measuring the gain of each model in terms of fulfilled agreements when compared to the baseline NT model, we verified that clients of type All outperformed S clients in 10%.

In the same way, even if the captured differences between results might be not too expressive, we verified that clients of type All outperformed clients SB ($t(29) = 1.88, p = 0.035$) and SI ($t(29) = 2.17, p = 0.019$), concerning variable F. The differences between these pairs of client agents were not significant, however, if using Bonferroni adjustments. The differences between the results obtained by clients of type All and clients of types S and SC were more expressive, in terms of variables F and V.

Discussion

When comparing the results obtained with the two alternative approaches to function $Tw_{x,y}$, we verified that the second approach lead to better results, in terms of variables F and V. In fact, when comparing the gain obtained in terms of outcomes of type F and V by each computational trust model relative to the trust-less NT model, we verified that this gain was more relevant

using the second approach, particularly in the cases of model `SC`, but also for models `S`, `SI` and `All`. Clients of type `SB` had similar results using both approaches to the trustworthiness evaluation function.

As concluding remarks, these experiments evaluating the integration of the trust-based components that we proposed in this thesis lead to promissory results about the utility of these components and the way of combining them. We are, however, aware that the proposed social-based integration of these components needs further work, and that other types of experiments need to be done in order to better evaluate this integration.

5.8 Concluding Remarks

In this chapter, we evaluated our approach to computational trust, more concretely, our current instantiation of the SOLUM framework, through experimental simulation. We started the chapter with a description of the generic selection process that served as the base scenario for our experiments, and we also described the methodology followed in the experimental analysis.

Then, we presented a brief phase of experiments with *Sinalpha*, and proceeded to the evaluation of the *Contextual Fitness* component. Here, we subdivided Hypothesis 2 into more fined tuned hypotheses that were tested in specific sets of experiments. We were able to conclude that the use of proper computational techniques that enable to extract contextual information from the set of evidence on the trustee under evaluation in fact improved the reliability of the estimation of this trustee's trustworthiness; that the consequent reliability of the trust decision was improved even when the available evidence was scarce; that *Contextual Fitness* performed better than an alternative approach for computational situation-aware trust based on pre-defined measures of similarity between situations; that *Contextual Fitness* supported the search for more desirable partners in mutualistic terms without jeopardizing the trust-based selection decisions; and that the benefits of *Contextual Fitness* were equally evident when this component was used in conjunction with different other types of trustworthiness estimators besides *Sinalpha*, namely, the well known proposals of Jøsang and Ismail (2002) and Jonker and Treur (1999).

In Section 5.4, we described a model of agents' behavior where decisions related to the enactment of agreements were made based on the ability and dispositional benevolence of the partners, and also on their mutualistic interests. With such a model, the agents were able to evolve their behavior with time, situation, and, more important, with the stage of the relationships they maintained with each one of their partners. Hence, this model of agents'

behavior played a key role in the evaluation of the *Integrity Tuner* and *Social Tuner* components, as described in subsequent sections. In fact, it is our proposal to address Research Question 4.

The evaluation of *Integrity Tuner*, which partially addressed Research Question 1 and was used to test Hypothesis 3, was presented in Section 5.5. From the results obtained, we were able to conclude that trusters that added the *Integrity Tuner* component to the trust-based evidence aggregator were able to make more reliable trust decisions than the ones that did not use this component. This was valid for all three evidence aggregators that we considered in the experiments. In the same way, the improved reliability in trustworthiness estimation was observed even when the populations of agents used in the experiments did not present specific characteristics concerning their dispositional integrity. Hence, we accepted the trustfulness of Hypothesis 3, in the conditions of the experiments.

The evaluation of *Social Tuner*, which addressed research questions 1 (partially) and 3, and that was used to test Hypothesis 1, was presented in Section 5.6.

From the results obtained, we were able to conclude that trusters that used *Social Tuner* in these circumstances made more reliable trust decisions than trusters that did not use this component; that the benefits of *Social Tuner* were equally evident when this component was used in conjunction with different types of trustworthiness estimators; and that trusters that used the *Social Tuner* component in environments not driven by benevolence-based behaviors did not made worse trust-based decisions than those not using this component. Hence, we accepted the trustfulness of Hypothesis 1, in the conditions of the experiments.

Finally, we evaluated the two approaches to integrate all developed trust-based components that we described in the previous chapter. Although these approaches reflect work in progress, the results we obtained and presented in Section 5.7 allowed us to conclude the usefulness of considering the ascription of different weights to the estimated values of ability, integrity and benevolence in different phases of the relationship between truster and trustee.

Chapter 6

Trust as a Service of the ANTE Platform

In the introductory chapter of this thesis, we questioned if the complementarity between trust and norms existing in our society could be transposed into computational artificial societies. Then, in Chapter 4, we formulated the hypothesis that the use of trust in agent-based systems for contracting decreases the weight of sanctions without jeopardizing the efficiency of normative control in promoting the compliance of agents. In order to evaluate this hypothesis, we integrated different components of our computational trust model in ANTE, the Agreement Negotiation in Normative and Trust-enabled Environments framework developed in the LIACC Laboratory in the scope of FCT project PTDC/EIA-EIA/104420/2008. This framework provides different services that assist the establishment, monitoring and enactment of electronic agreements in a semi-automatic way.

In this chapter, we present the work conducted to test the hypothesis mentioned above. The research performed started with an exploratory phase of literature revision on computational trust and normative control, after which it was mostly empirical, experimental, and simulation-based. The settlement of the experimental testbed included the accommodation of our computational trust model into ANTE, as an autonomous service of this framework, and its integration with two other fundamental services of the framework: the negotiation facilitator and the normative environment.[1]

[1]The work presented in this section is the result of the joint work of the LIACC members contributing to the ANTE platform. Some of the text that appears in this section was adapted from (Urbano et al., 2010a, 2011a; Lopes Cardoso et al., 2012, 2013).

6.1 Introduction

The ANTE platform is a multi-agent system that assists the automatic negotiation of agreements between heterogeneous agents, by regulating the social behavior of these agents through the use of explicit and implicit rules and norms. It addresses the issue of multi-agent collective work in a comprehensive way, covering both negotiation as a mechanism for finding mutually acceptable agreements, and the enactment of such agreements. Furthermore, the framework also includes the evaluation of the enactment phase, with the aim of improving future negotiations. In a sense, ANTE follows and expands the notion of electronic institution as a means for delivering regulated multi-agent environments given in (Lopes Cardoso and Oliveira, 2005; Lopes Cardoso, 2010).

Taking a broad perspective, an agreement in this context is a solution obtained using a distributed cooperative problem solving approach. Therefore, a wide range of problems can be tackled. The agreement binds each negotiation participant to its contribution to the overall solution. It is therefore useful to represent the outcome of a successful negotiation process in a way that allows for checking if the contributions of each participant do in fact contribute to a successful execution of the agreement. A normative environment, within which agent interactions that are needed to enact the agreement will take place, takes care of this monitoring stage. Assessing the performance of each contribution is essential to enhance future negotiations. Computational trust may therefore be used to appropriately capture the trustworthiness of negotiation participants, both in terms of the quality of their proposals when building the solution (i.e. the practicability of the approach) and in terms of their ability to successfully enact their share.

6.1.1 Services of the ANTE Framework

The ANTE framework provides the following services:

- Negotiation facilitator (*nfService*): provides assistance in using specific negotiation protocols.
- Ontology mapping (*omService*): provides ontology mapping services for users that use their own disparate domain ontologies.
- Notary (*nService*): trusted third party that registers digitally-signed contracts.
- Normative environment (*neService*): is responsible to check whether a set of interacting users behave according to a set of applicable norms.

Contracts specify the norms contractual users commit to, and these norms are embedded into the normative environment, which therefore provides a contract monitoring and enforcement facility.

- Computational trust (*ctService*): provides trustworthiness estimations of agents of the platform. It constitutes a social mechanism for discouraging contractual deviations. This service integrates different components of our computational trust model, described in Chapter 4.

- Other services related to third-party entities, such as banks and delivery tracking entities.

The current version of the ANTE platform is an agent-based system implemented using Jade technology (Bellifemine et al., 2007). Hence, we assume that each different service of ANTE is managed by a service agent, even if this type of agents may lack some of the autonomic characteristics normally referred in the common agent definition. There are different types of service agents; however, taking into consideration that our work in this thesis focus on the interactions between the *ctService*, *neService*, and *nfService*, we only consider agents of type *ctServiceAgent*, *neServiceAgent*, and *nfServiceAgent*.

In the same way, we consider the existence of user agents representing real users registered in the platform seeking for assistance in sourcing, establishing and enacting their agreements. This way, client agents (*clientAgent*) represent users that, at any given moment, identify a new opportunity to interact and start a process to select the best partners to interact with in the scope of this opportunity. In turn, provider agents (*providerAgent*) represent users that respond to the clients' calls for participation. It is worth to note that every user of the platform can have both roles at any given moment. There is a third type of user agents, the administrators of the ANTE platform, which we do not consider in the context of this work.

Finally, the platform also considers third party agents representing entities such as banks, delivery track systems and messengers. However, as these entities do not directly interfere with the interconnection between services *ctService*, *neService*, and *nfService*, we do not characterize these agents any further in this thesis.

6.1.2 Trust-based Establishment of Agreements

The generic process of establishing electronic agreements between user agents in ANTE is illustrated in the UML-based sequence diagram of Figure 6.1.

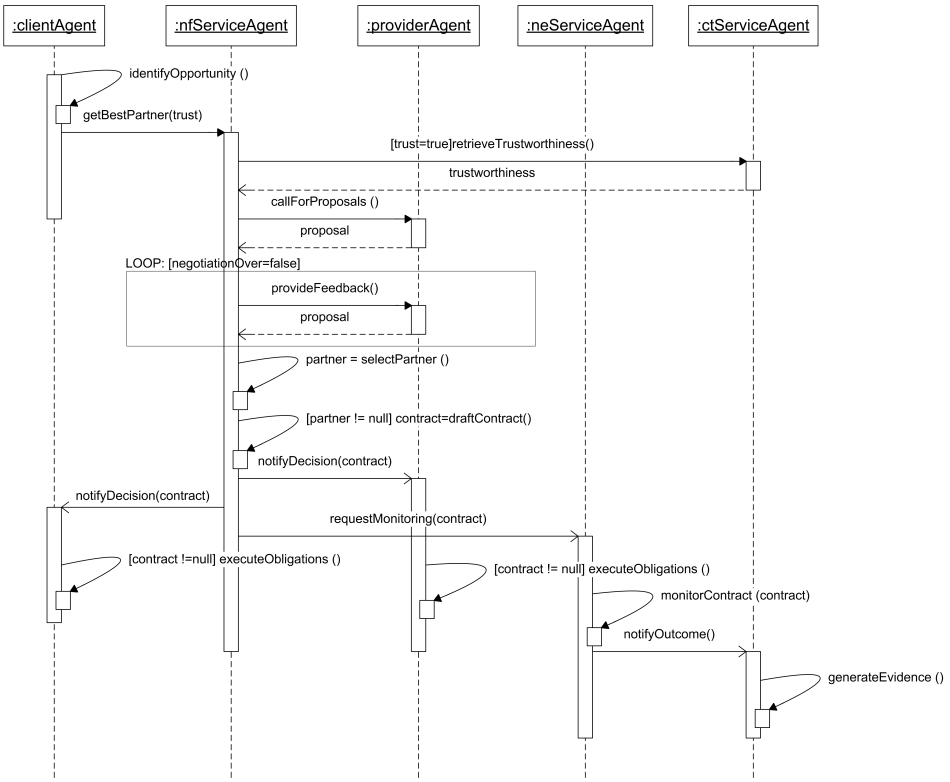

Figure 6.1: Sequence diagram for trust-based establishment of agreements in ANTE.

We divide this process in two main phases: selection of partners (including contract drafting), and monitoring/generation of evidence.

Selection of Partners

The phase of selection of partners starts when a client agent (on behalf of the user it represents) identifies a new opportunity for interaction (e.g., a business opportunity). It then requests assistance to the *nfService* (through an *nfServiceAgent*) to find the best partner to interact with within the scope of the identified opportunity. It is up to the *nfServiceAgent* to search all possible provider agents that fit to the current opportunity, using, for example, a directory service provided by the ANTE platform (this step is not shown in the sequence diagram). Then, the *nfServiceAgent* asks the *ctServiceAgent* to compute and send the trustworthiness of these possible provider agents. After this step, the *nfServiceAgent* sends a call for proposals to the selected provider agents concerning the exploration of the opportunity to interact. The selected providers respond by proposing their own terms concerning this

interaction. The call for proposals initiates the negotiation process between the *nfServiceAgent* and the candidate provider agents, following the built-in multi-round, multi-attribute, feedback-based negotiation protocol described in (Rocha and Oliveira, 1999) provided by the *nfService* service. If the client agent showed its intention to perform a trust-based selection of partners, the trustworthiness of the candidate partners is then used as another dimension of the negotiation protocol (we address the negotiation attributes later in this chapter). At the end of the negotiation process, the *nfServiceAgent* selects the best partner to interact with the client agent it represents, drafts a contract defining the terms of the agreement, and notifies all provider agents of the outcome of the negotiation process. The selected agent (if any) receives the contract just drafted. The client agent also receives this contract, which includes the name of the selected provider agent (if any). If the negotiation process completes without a selected agent, the process of agreement establishment terminates without success.

Monitoring/Generation of Evidence

After the selected provider is notified about the negotiation outcome and both partners sign the contract (not shown in the diagram), the *nfServiceAgent* requests the *neServiceAgent* to monitor the execution of this contract. Finally, when the contract is enacted, the *neServiceAgent* sends specific information about the contract's execution to the *ctServiceAgent*, which generates a new piece of evidence from this information that may be used in later assessments of both partners' trustworthiness.

6.1.3 More About the *ctService*

The current version of ANTE provides a built-in trust function that uses trust-based information generated within the platform from the activity of registered users. In future versions of the platform, we intend to provide an API in order to allow user agents to consider other trust-based information obtained by their own means. In the same way, we intend to allow user agents to use their own trust formulas.

We must also refer that the current stabilized version of the *ctService* service implements a much simpler version of the SOLUM model than the one presented in Chapter 4. In reality, the trustworthiness estimation is based on the *Sinalpha* component and, depending on the user option, on *Contextual Fitness*. The formulas of the trustworthiness evaluation function in both cases (i.e., considering situation-less or situation-aware assessments) are the ones considered in equations 5.4 and 5.5, respectively. It is worth to note,

however, that the most updated version of SOLUM is being implemented in ANTE as of the writing of this thesis.

In the next section, we present a more detailed view of the role of the *ctService* in ANTE and the way it interacts with two other key services of ANTE, *neService* and *nfService*.

6.2 The Role of *ctService* in ANTE

In Figure 6.1, we showed the basic interactions between user agents and the key service agents of ANTE. In this section, we provide more detail on specific parts of the process of establishing agreements where computational trust plays an important role.

6.2.1 Trust-based Pre-Selection of Partners

We added to ANTE the possibility of preselecting the provider agents by their trustworthiness, at the phase of selection of partners. This way, if the client agent showed its intention to use this type of pre-selection, the *nfServiceAgent* selects the $n > 0$ most trustworthy provider agents to which it will send a call for proposals, just after retrieving the trustworthiness of the candidate provider agents from the *ctServiceAgent*. All other provider agents are automatically excluded from negotiation. This pre-selection may be useful in several real scenarios. For example, a tech firm that is announcing a position for Java programmers for which it has received more than three hundred applications may want to have the possibility of preselecting the best candidates according to their perceived trustworthiness (derived from potentially diverse information, such as the one resulting from curricula analysis and third party opinions), before pursuing to a deeper and more expensive analysis of the candidates.

Figure 6.2 shows the window where a given registered user of ANTE wanting to start a negotiation configures the negotiation parameters. At the left panel, the client specifies the elements that shall be present in the call for proposals. In the right panel, the client has the option to specify different trust-based parameters related to the negotiation protocol. If he wants to preselect partners, he sets the `Top N` form field in the top panel to the number of providers that shall enter the negotiation. Currently, this field has half a dozen of positive values (e.g., 5, 10, 20, ...). Then, the client must confirm the use of pre-selection by ticking the correspondent box in the bottom panel. Additionally, the client may indicate that he is looking for situation-aware trustworthiness estimation, ticking the `Contextual`

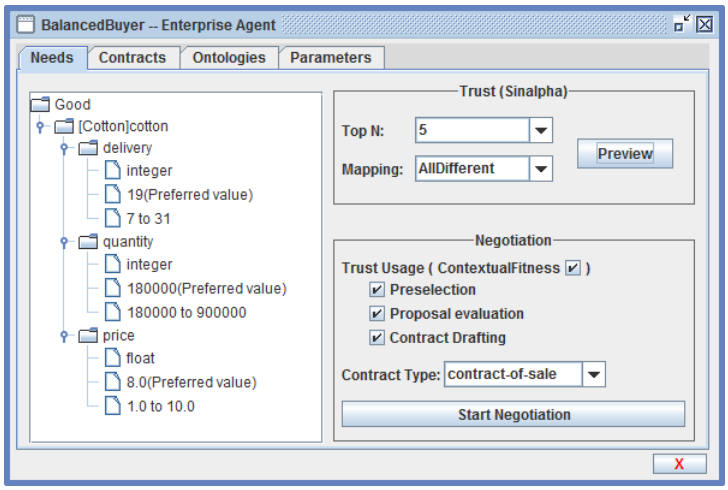

Figure 6.2: Configuration of a new negotiation by a client in ANTE.

`Fitness` option. He also has the possibility of previewing the characteristics of the N more trustworthy providers before starting the negotiation. At last, when the client presses that button `Start Negotiation`, the client agent representing the user asks the *nfServiceAgent* to start the negotiation phase, as described before.

6.2.2 Trust-based Proposal Evaluation

As we have mentioned before, clients may choose to use trust in the evaluation of the proposals sent by the providers in the negotiation phase. For this, they must tick the option `Proposal evaluation` in panel `Negotiation` (see Figure 6.2).

The current version of ANTE uses the Q-negotiation protocol (Rocha et al., 2005) to support the process of selection of partners. Using this protocol, negotiation participants engage themselves in a sequential negotiation process composed of multiple rounds, by exchanging multi-attribute proposals and counter-proposals, trying to convince each other to modify the values for attributes they evaluate the most. At every negotiation round, the Q-negotiation protocol chooses the proposal that is estimated to make the best deal. When using trust in proposal evaluation, this decision is based on the estimated utility of each providers' proposal and on these providers' estimated trustworthiness, as shown in Equation 6.1. In the equation, $\omega_{tr} \in [0, 1]$ is a weighting parameter that allows to configure the importance assigned to the trust component in this selection method, and $\mathcal{A}'_{providers}$ is the set of all candidate partners to this negotiation.

$$D_x^{tr-up} = \arg \max_{y_i \in \mathcal{A'}_{providers}} [\omega_{tr} \cdot tr_{x,y_i} + (1 - \omega_{tr}) \cdot up_{y_i}] \ . \tag{6.1}$$

The utility of a given proposal up_{y_i} as measured by client x is related to how close the conditions offered by provider y_i are to the terms of the cfp specified by x, as formalized in Equation 6.2 (adapted from Rocha and Oliveira, 1999), where $v_{i,p_{y_i}}$ is the value of the negotiation attribute i of proposal p_{y_i}, k is the number of negotiation attributes considered, and $v_{i,min}$ and $v_{i,max}$ define the domain of possible values to be negotiated by attribute.

$$up_{y_i} = 1 - \frac{1}{k} \times \sum_{i}^{k} \frac{|v_{i,cfp} - v_{i,p_{y_i}}|}{v_{i,max} - v_{i,min}} \ . \tag{6.2}$$

As results evident, if the client does not select the `Proposal evaluation` option in panel `Negotiation`, the selection of the best proposal at any round is made considering that $D_x^{up} = \arg\max_{y_i \in \mathcal{A'}_{providers}}(up_{y_i})$.

6.2.3 Trust-based Drafting of Contracts

In ANTE, a contract is a set of contractual obligations, such as the delivery and payment obligations; it defines a normative relation of a given type within which a group of agents commits to a joint activity. Moreover, it includes a set of contractual information that makes up a kind of background knowledge for that contract. The use of contracts in ANTE is described in (Lopes Cardoso, 2010).

When configuring a negotiation (Figure 6.2), the client can tick the `Contract Drafting` option in the `Negotiation` panel. Our idea is to allow the automatic drafting of the contractual terms of the agreement, specifying more or less obligations and/or imposing heavier or lighter sanctions, taking into account the estimated trustworthiness of the partner agents. Therefore, trustworthy agents could benefit from lighter contracts, while the interactions with less trustworthy agents could be somewhat protected by the addition of obligations and sanctions to the letter of the law.

Currently, our work on trust-based drafting of contracts is restricted to the automatic choice of sanctions to be applied to any given contract, given the estimated trustworthiness of partners. A detailed description of our current empirical work on this topic is given in Section 6.3.2.

6.2.4 Generation of Trust-based Evidence

The ANTE platform allows to generate evidence about the behavior of registered agents from the monitoring service of *neService*. This evidence respects

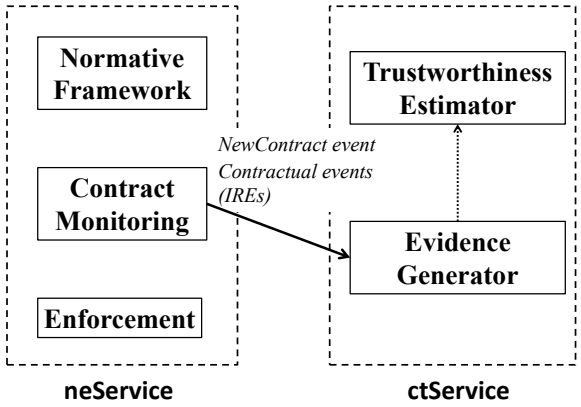

Figure 6.3: Asynchronous notification of contractual events from *neService* to *ctService*.

the format defined in Section 4.2.4, and can be used by the *ctService* service to feed the trust-based evaluation functions. In terms of agents, the *ctServiceAgent* subscribes the monitoring service of *neServiceAgent* at setup. Consequently, the *ctServiceAgent* service starts receiving asynchronous reports about the contractual events that result from the monitoring activity for all contracts that are established in ANTE. Figure 6.3 illustrates this process.

We use the model of contractual obligations proposed by (Lopes Cardoso and Oliveira, 2010), which allows for a rich set of possible contract enactment. Using this model, the monitoring service of the normative environment launches different occurrences regarding the contract being monitored, which are named *institutional reality elements* (*IRE*). Different types of *IRE* may be distinguished, as follows:

$StartContract^C(t)$	contract C has started at time t
$Ifact^C(f)^t$	fact f is recognized as having occurred at time t
$Time^C(t)$	instant t has elapsed
$Obl^C_{b,c}(f \prec d)$	agent b is obliged towards agent c to bring about f until d
$DViol^C(obl)^t$	there was a deadline violation of obligation obl at time t
$Fulf^C(obl)^t$	obligation obl was fulfilled at time t
$Viol^C(obl)^t$	obligation obl was violated at time t
$EndContract^C(t)$	contract C has ended at time t

Also, in the scope of this work, we consider the existence of one specific contract, *the contract of sale*, which stipulates an obligation of *Delivery* from provider y toward client x and an obligation of *Payment* from client

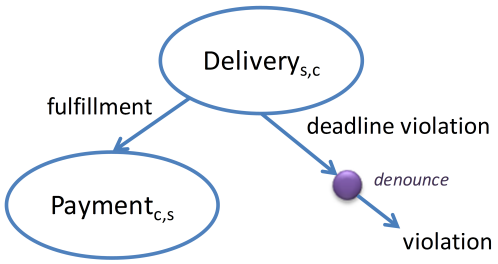

Figure 6.4: The simple *contract of sale*

x toward provider y. The contract of sale is depicted in Figure 6.4. When this contract is instantiated in the ANTE platform, two different contractual evidences are generated: $e_{x,y}$, which represents the behavior of the provider concerning the delivery action, and $e_{y,x}$, which represents the behavior of the client concerning the payment to the provider. For the sake of clarity, we describe next the generation of evidence $e_{x,y}$.

Generation of Evidence $e_{x,y}$

The evidence $e_{x,y}$ is generated by *ctServiceAgent* from the *IREs* sent by the *neService*. Values $v_1^{e_{x,y}}$ and $v_2^{e_{x,y}}$ (identity of the contractual parties) and $v_5^{e_{x,y}}$, $v_6^{e_{x,y}}$ and $v_7^{e_{x,y}}$ (task activity) are derived automatically from the $StartContract^C(t)$ *IREs*, in a straightforward manner. Value $v_3^{e_{x,y}}$ (time) is generated from the time t of reception of the $EndContract^C(t)$ *IRE*. The final value, i.e., the outcome of the evidence ($v_8^{e_{x,y}}$), is generated from the *IRE*s received by *ctServiceAgent* after $StartContract^C(t)$ and before $EndContract^C(t)$ are received. The procedure for the generation of $v_8^{e_{x,y}}$ is further divided in the following steps:

1. Determine set $\mathcal{O}$ from the type of contract. In the case of the contract of sale, $\mathcal{O} = \{F, Fd, V\}$.

2. Map the received *IRE*s into $\mathcal{O}$. In the case of the contract of sale, we make $\{Obl_{b,c}^C(f \prec d), Fulf^C(obl)^t\} \rightarrow F$, $\{Obl_{b,c}^C(f \prec d), DViol^C(obl)^t, Fulf^C(obl)^t\} \rightarrow Fd$, and $\{Obl_{b,c}^C(f \prec d), DViol^C(obl)^t, Viol^C(obl)^t\} \rightarrow V$.

6.2.5 Interface to *ctService*

The ANTE platform offers an interface to the computational trust service, shown in Figure 6.5, which allows the inspection of how trustworthiness as-

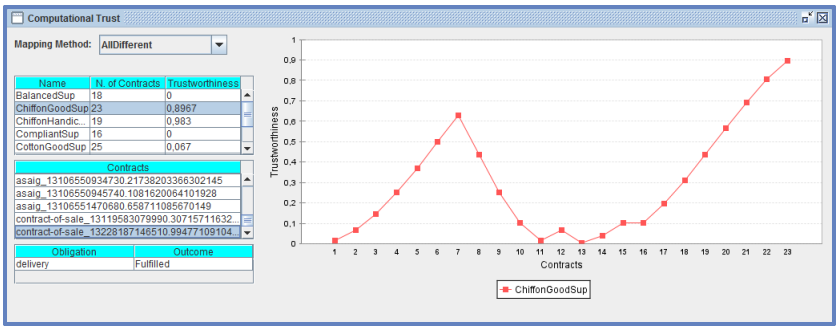

Figure 6.5: Computational trust: computing trustworthiness assessments from contractual evidences.

sessments are being computed, including the visualization of the contractual evidences that are used as input for each evaluated agent.

6.3 Experimental Studies in the ANTE Platform

The ANTE platform is being developed to assist the development of different agreement technologies. It also allows us to conduct interesting studies about the interface between negotiation services, normative environment services, and computational trust services. In this section, we describe two of such studies that we have conducted in the ANTE platform. The first evaluates the benefits that may exist when using computational trust to enhance the outcome of automatic negotiations. We present this study in Section 6.3.1. The second study is related to Hypothesis 4 that we formulated in Chapter 4, where we hypothesized that the use of trust in contracting agent systems decreases the weight of sanctions without jeopardizing the efficiency of normative control in promoting the compliance of agents. We present this study in Section 6.3.2.

6.3.1 Trust at Different Negotiation Stages

The majority of the papers on computational trust assumes that trust is the only dimension to take into attention when selecting partners. The works of (Castelfranchi et al., 2003; Maximilien and Singh, 2005; Kerschbaum et al., 2006) refer that trust must be used additionally to other relevant dimensions, but do not provide a practical study on the complementary use of such dimensions. Gujral et al. (2006) and Griffiths (2005) propose models of partner selection based on multi-dimensional trust but do not refer the pre-selection phase. The work by Padovan et al. (2002) develops a sce-

nario that depicts a small value chain, where the selection of partners is performed by ranking the received offers by the assessed offer price, which includes the expected value of loss based on a reputation coefficient. However, this work does not consider pre-selection. The work by (Kerschbaum et al., 2006) addresses the problem of member selection in virtual organizations and considers the possibility of selection of candidate partners based on the reputation of agents, prior to the negotiation phase. The authors also consider the use of trust in the negotiation phase, both as another negotiation dimension, such as price and delivery time, or as a factor in deciding between equally well-suited candidates. However, the empirical evaluation of their trust model is focused on testing its resistance to attacks, and they do not model negotiation in their experiments.

Based on this, we conducted an experimental analysis on the effect of using different trust-based selection methods – including pre-selection and the use of trust in the negotiation phase – on the utility of the selecting agents.

Motivation

There may exist specific real-world situations where the most trustworthy agent is not the one that offers the best payoff to the selecting agent. Let us consider two hypothetical examples. In the first example, firm A is a manufacturer of t-shirts and firms B and C are providers of fabric. Firm A knows, from experience, that B rarely fails a contract. In the same way, it also knows that C is less reliable, and sometimes it delays a delivery; however, firm C offers better utility (possibly derived from better quality of the product or better shipment and payment conditions) than firm B when it does not breach the contracts. In this case, the fact that B is more trustworthy than C can mean that B is more useful to A than firm C? The second example depicts a recruitment scenario and is related to the use of trust in the selection decision as a *pre-filtering* activity. In the example, firm D has one position open for Java programmers for which it has received more than three hundred applications. The firm has the possibility of preselecting the best candidates according to their trustworthiness, before pursuing to a deeper and more expensive analysis of the candidates. In this case, how many candidates shall be returned by the filtering process? In this section, we address the questions raised in the examples above. In particular, we study the effect of using different methods based on trust for selecting partners. This study is enhanced by considering two distinct situations: in the first one, the proposals received by a client in a negotiation process are

Table 6.1: Configuration parameters (trust in different negotiation stages).

terms of the cfp	randomly assigned at each round
$\mathcal{V}_5$	*cotton*, *chiffon*, *voile*
$\mathcal{V}_6$ (quantity)	$\{q \in \mathbb{N} : q \in [v_{quant,min}, v_{quant,max}]\}$
$\mathcal{V}'_6$ (price)	$\{p \in \mathbb{N} : p \in [v_{price,min}, v_{price,max}]\}$
$\mathcal{V}_7$	$\{d \in \mathbb{N} : d \in [v_{dtime,min}, v_{dtime,max}]\}$
$\mathcal{O}$	F, V
terms of proposals	randomly assigned at each round
selection criteria	$D^{tr}_{x,y}$
# client agents	20
# provider agents	50
# rounds	30
# experiment repetitions	20
Sinalpha ω, λ_F and λ_V	$\pi/12$, 1.0 and -2.0
Sinalpha $lf(F)$ and $lf(V)$	0 and 1

relatively similar and yield comparable utility to the client. In the second situation, the proposals are more disparate.

Scenario and Notation

We use the generic selection process and the methodology described in sections 5.1.1 and 5.1.2, respectively. We instantiate this process to the textile importer-exporter transactions domain. Also, the generic configuration parameters of the experiments is shown in Table 6.1. The values $v_{i,min}$ and $v_{i,max}$ define the minimum and maximum values allowed for attribute i, respectively. The terms specified by providers in their proposals are generated randomly following a uniform distribution in the range $[v_{i,p,min}, v_{i,p,max}]$, where $v_{i,p,min}$ and $v_{i,p,max}$ are defined in equations 6.3 and 6.4, respectively. Also, $v_{i,cfp}$ is the value defined in the *cfp* for attribute i (quantity, price or delivery time), and $\delta \in [0, 1]$ is a *dispersion* parameter that allows to define how distant the generated proposal is from the preferences of the client, as stated in the *cfp*.

$$v_{i,p,min} = \max\ ((1 - \delta) \times v_{i,cfp},\ v_{i,min})\ . \tag{6.3}$$

$$v_{i,p,max} = \min\ ((1 + \delta) \times v_{i,cfp},\ v_{i,max})\ . \tag{6.4}$$

After calculating the utilities of all received proposals, the client makes a decision concerning the selection of the *best* proposal. In this paper, we analyze three different approaches for the selection of the best proposal:

Table 6.2: Different types of experiments, based on the places where trust was used

#	Selection Method	Pre-selection	Negotiation
1	No Trust	—	—
2	Trust in pre-selection (10%)	√	—
3	Trust in pre-selection (50%)	√	—
4	Trust in negotiation	—	√
5	Trust in pre-selection (10%) and in negotiation	√	√
6	Trust in pre-selection (50%) and in negotiation	√	√

1. Proposals are sorted by their utility (as calculated in Equation 6.2), and the best proposal is the one that has the highest utility ($D^{up}_{x,y}$).

2. Proposals are sorted by the trustworthiness of the proponent providers, and the best proposal is the one which corresponds to the highest value of trustworthiness ($D^{tr}_{x,y}$).

3. Proposals are sorted by the weighted sum of their utility and the trustworthiness of the corresponding proponents (cf. Equation 6.1), and the best proposal is the one that presents the highest value for this weighted sum.

In all methods using trust estimation, the *Sinalpha* and *Contextual Fitness* components are used. The resulting trustworthiness evaluation function is shown in Equation 5.5. Also, the behavior of providers is defined by their handicaps, and each provider is randomly assigned a handicap following a uniform distribution over all possible handicaps presented in Table 5.1.

Experiments

We ran six different experiments, according to the selection methods under evaluation. Table 6.2 presents these experiments. As can be observed in Table 6.2, we tested two different filtering approaches (experiments 2, 3, 5 and 6): the first one preselected 10% of the most trustworthy providers registered in the ANTE platform, and the second one preselected 50% of this population. In experiments 1 and 4, no trust-based pre-selection was performed and all providers were allowed to proceed to the negotiation phase.

In order to enhance our study on the effect of using trust in selection processes, we considered two different values for the dispersion parameter δ: 0.2 and 1.0 (cf. equations 6.3 and 6.4). As mentioned before, parameter δ is used to configure how distant the proposals generated by the providers are from the conditions specified in the received *cfp*. In these experiments, the value 0.2 was used to configure small deviations, which means that all the

proposals received by the client agent were close to its preferential values for current interaction; in opposition, the value of 1.0 allowed for a greater dispersion in the utility of the proposals received by the client agent.

Evaluation Metrics

In order to evaluate and compare each one of the selection methods considered in the experiments, we used six different performance metrics. The first metric was the *utility of the interaction* (μ_t), given in Equation 6.5. We averaged this utility over all clients and all episodes.

$$\mu_t = \begin{cases} up, & \text{if } o = f \ , \\ 0, & \text{if } o = v \ . \end{cases} \tag{6.5}$$

The second metric was the *number of positive outcomes* (o^+) obtained by all client agents in an episode, averaged over all episodes. The third metric was the *number of different providers* (Δ_{sup}) selected by all clients in one episode, averaged over all episodes. The fourth and the fifth metrics measured the *trustworthiness of the provider* and the *utility of the proposal* selected by a client in one episode (τ_s and μ_s, respectively), averaged over all clients and all episodes. Finally, the sixth metric was the *number of unfitted choices* (ζ) performed by a client, averaged over all clients and all episodes. This latter metric is related to the *Contextual Fitness* component of our computational trust model. It concerns the choice of a provider that the client knows has an handicap in the current business conditions.

Results with $\delta = 0.2$.

The first part of the experiments was performed using $\delta = 0.2$. We first measured the average utility of the proposals *received* by a client in one episode and averaged it over all clients and all episodes. The value we obtained for this average was 0.93, with a standard deviation of 0.03. These values were obtained consistently for all the selection methods tested. Their meaning is that the providers offered proposals with approximated utility and close to the clients' preferences.

Table 6.3 presents the results obtained in this first set of experiments for the defined metrics. In experiments 4.x, 5.x and 6.x, the utility of the interaction (μ_t) is a weighted sum of the trustworthiness of the provider and the utility of its proposal. In the experiments, we used two different values for the weight of the trust component, $\omega_\tau = 0.1$ and $\omega_\tau = 0.5$).

From the results presented in Table 6.3, we verify that the selection method that did not rely on trust got worse results for the metric utility of

the interaction ($\mu_t = 0.69$), as it would be expected. This method selected the providers by the utility of their proposals, which allowed for the selection of proposals with very high values of utility ($\mu_s = 0.98$) and for a high degree of exploration of new partners ($\Delta_{sup} = 0.84$). However, the trustworthiness of the selected providers was in average very low ($\tau_s = 0.17$), and a relevant number of unfitted choices was done ($\zeta = 0.21$). In consequence, the number of positive outcomes was relatively low ($o^+ = 0.70$).

The results presented in Table 6.3 also show that the mixed use of trust, both in pre-selection and in the negotiation phase (experiments 5.x and 6.x), got the best results in terms of the utility of interaction ($\mu_t \approx 0.83$), for all combinations of the degree of filtering (10% and 50%) and ω_τ. In this case, we verified that although reinforcing the trust component in the negotiation phase ($\omega_\tau = 0.5$) allowed for higher values of the trustworthiness of the selected providers (τ_s), relaxing this value ($\omega_\tau = 0.1$) allowed for higher values of the utility of the selected proposals (μ_s). Also, the difference between filtering the 10% or the 50% more trustworthy agents was not relevant for $\delta = 0.2$. Finally, we observed that both the use of standalone, stricter pre-selection (10%) and the use of trust in negotiation with $\omega = 0.1$ allowed for similar good results of μ_t (0.82), and approximated values of o^+, τ_s and μ_s. The use of standalone, more relaxed pre-selection (50%) and the use of trust in negotiation with $\omega = 0.5$ got lower values of μ_t (0.79), with the first method exploring more the utility of the proposals ($\mu_s = 0.98$) in detriment to the trustworthiness of providers ($\tau_s = 0.41$), and the latter having an opposite behavior ($\mu_s = 0.93$ and $\tau_s = 0.90$).

Experiments with $\delta = 1.0$.

In the second part of the experiments, we wanted to evaluate the effect of each one of the selection methods when the dispersion in the utilities provided by different providers was bigger. For that, we configured δ to have value 1.0. In this case, the measured value for the average utility of the received proposals was 0.73, with a standard deviation of 0.11, showing a higher variance in the proposals made by the providers.

Table 6.4 presents the results obtained in this second set of experiments for the metrics described before.

Table 6.3: Results obtained with $\delta = 0.2$

#	Selection Method	μ_t	o^+	Δ_{sup}	τ_s	μ_s	ζ
1	No Trust	0.69	0.70	0.84	0.17	0.98	0.21
2	Trust in pre-selection (10%)	0.82	0.85	0.35	0.80	0.96	0.00
3	Trust in pre-selection (50%)	0.79	0.81	0.75	0.41	0.98	0.01
4.1	Trust in negotiation ($\omega_\tau = 0.1$)	0.82	0.87	0.23	0.83	0.95	0.00
4.2	Trust in negotiation ($\omega_\tau = 0.5$)	0.79	0.85	0.11	0.90	0.93	0.00
5.1	Trust in presel. (10%) & in neg. ($\omega_\tau = 0.1$)	0.83	0.88	0.18	0.88	0.95	0.00
5.2	Trust in presel. (10%) & in neg. ($\omega_\tau = 0.5$)	0.82	0.88	0.11	0.90	0.93	0.00
6.1	Trust in presel. (50%) & in neg. ($\omega_\tau = 0.1$)	0.83	0.87	0.22	0.85	0.95	0.00
6.2	Trust in presel. (50%) & in neg. ($\omega_\tau = 0.5$)	0.83	0.89	0.11	0.91	0.93	0.00

Table 6.4: Results obtained with $\delta = 1.0$

#	Selection Method	μ_t	o^+	Δ_{sup}	τ_s	μ_s	ζ
1	No Trust	0.66	0.71	0.83	0.17	0.93	0.21
2	Trust in pre-selection (10%)	0.73	0.87	0.36	0.80	0.84	0.00
3	Trust in pre-selection (50%)	0.73	0.80	0.76	0.41	0.92	0.02
4.1	Trust in negotiation ($\omega_\tau = 0.1$)	0.75	0.83	0.63	0.58	0.91	0.00
4.2	Trust in negotiation ($\omega_\tau = 0.5$)	0.67	0.88	0.14	0.88	0.77	0.00
5.1	Trust in presel. (10%) & in neg ($\omega_\tau = 0.1$)	0.73	0.87	0.32	0.83	0.85	0.00
5.2	Trust in presel. (10%) & in neg ($\omega_\tau = 0.5$)	0.66	0.86	0.13	0.89	0.77	0.00
6.1	Trust in presel. (50%) & in neg ($\omega_\tau = 0.1$)	0.77	0.85	0.59	0.64	0.90	0.00
6.2	Trust in presel. (50%) & in neg ($\omega_\tau = 0.5$)	0.66	0.86	0.14	0.89	0.77	0.00

The results obtained and presented in Table 6.4 show relevant differences from the results obtained with $\delta = 0.2$. In fact, the combined use of trust in pre-selection and in negotiation did not achieve the same good performance as observed with $\delta = 0.2$, for $\omega_\tau = 0.5$. As illustrated in Table 6.4, in experiments 5.2 and 6.2, the clients kept selecting the same trustworthy agents again and again ($\Delta_{sup} \approx 0.14$), showing a rather parochial behavior. This had the cost of decreasing the utility of the selected proposals ($\mu_s = 0.77$) in a significant manner, with just a slight improvement in the trustworthiness of the selected providers ($\tau_s = 0.89$). In a general case, we can observe in Table 6.4 that all trust methods that used trust in negotiation with a strong weight for the trust component ($\omega_\tau = 0.5$) got as little value for μ_t as the selection approach that did not use trust at all. In the same way, approaches using more restricted pre-selection (10%) exhibited significantly lower values of μ_t than their counterparts using $\delta = 0.2$.

The results obtained also show that the combined use of a more relaxed filtering of providers (50%) and a lower weight of the trust component ($\omega_\tau = 0.1$) had again achieved the best result for the average utility of interaction ($\mu_t = 0.77$). This approach allowed for a better equilibrium between the trustworthiness of the selected providers and the utility of the selected proposals.

Interpretation of Results

The results obtained and presented in the sections above allow us to conclude that parochialism in partner selection is acceptable when the proposals under evaluation are not too disparate ($\delta = 0.2$). In this case, selection methods strongly supported on trust revealed to be good choices, as they were able to select more reliable partners without loosing utility.

However, we have shown that when the standard deviation of the utility of the received proposals was about 11% of the mean, the excessive use of trust was not acceptable, as parochialism prevented clients from exploring partners that offered deals with higher utilities. In both the situations that we have studied, a method that preselects half of the population of candidate providers and then moderately uses trust in negotiation revealed to be a better choice (experiments 6.1 in tables 6.3 and 6.4).

6.3.2 Joint Use of Trust and Norms

Scenario and Agents Behavior Model

A truster starts by sending a call-for-proposals for a particular service, for which each trustee will provide its own proposal. When assessing proposals, trusters take into account their utility and (optionally) the perceived trustworthiness of each proponent. The truster will try to establish a contract with the proponent of the better proposal, for which it may decide to include control mechanisms in the contract. If, for some reason, the trustee is not able to accommodate this contract, the truster will try with the proponent of the second best proposal, and so on. At the contract enactment phase, each hired trustee will have the opportunity to fulfill the contract or to violate it, according to the behavior model described later.

Trusters

The truster behavior model is based on the interplay between trust and control, as discussed in Das and Teng (1998) and Tan and Thoen (2000). When considering the establishment of a contract with a trustee, the truster computes a confidence threshold Ct that indicates the minimum confidence he needs for entering into that particular transaction. This value is calculated by weighting the perceived risk R by the agent's risk aversion Ra. Risk, in turn, is modeled as a function of the weight of the transaction volume Tv on the agent's overall production volume Pv and the perceived trustworthiness T of the trustee, computed dynamically using a computational trust model. We thus have that $Ct = R * Ra$, where risk $R = Tv/Pv * (1 - T)$.[2] Risk aversion ranges from 0 (a risk lover agent) to 1 (totally risk averse).

Having a minimum confidence threshold, the truster will propose, to a selected trustee, a contract that includes a level of control (represented as a sanction to apply in case of violation) computed according to the general notion from Das and Teng (1998) that *Confidence = Trust + Control*. By suggesting an appropriate sanction, the truster tries to raise his confidence on the contract that is to be established with a particular trustee, of which it has some trustworthiness assessment.

[2] In the experiments, we use T/ζ instead of T, due to the fact that computational trust models typically overrate the trustworthiness estimations, as they tend to aggregate the outcomes of past evidence using statistical methods, without taking into consideration the *relationship* that was active between interacting partners at the evidence time (Urbano et al., 2011b).

Trustees

The model of behavior of the trustees is inspired in the model of betrayal in organizations of Elangovan and Shapiro (1998). In our model, trustees of low integrity tend to enter in new agreements even when they do not have enough resources to satisfy them, i.e. they are aware that they may have to (voluntarily) violate one or more of their active contracts. On the contrary, trustees with high integrity may refuse the contract if they do not have enough resources to satisfy the deal without violating previous agreements. We also assume that all trustees have a predefined level of *competence*, i.e. an innate ability to provide products of good quality. Violations that are due to (lack of) ability are not voluntary, and thus are not considered *betrayals* (Elangovan and Shapiro, 1998).

In the equations that follow, x denotes a trustee, y denotes a truster, c denotes a contract, and p denotes a contract proposal. When time-stamping terms using a superscript, we assume a discrete time line. Unless otherwise noted, variables are assumed to be universally quantified.

The decision to betray vs. keep the status quo is made in a process that starts when the trustee is selected for a new contract. The new contract is considered a business opportunity. As such, the agent will probabilistically consider betraying one of his active contracts. This betray propensity (henceforth ρ) is inversely proportional to the integrity of the trustee, denoted as $\delta \in [0, 1]$. We may say that with a probability ρ the trustee will consider betraying one of his active contracts c, provided that freeing the resources that are allocated for upholding that contract enables him to accept the new contract (i.e. when Equation 6.6 holds).

$$FreeResources(x) + Resources(c) > Resources(p) \ . \tag{6.6}$$

Active contracts are analyzed in decreasing order of utility. It is worth noting that even when the trustee already has enough free resources to encompass the new contract, it will still consider betraying one of his active contracts. After identifying a new opportunity, the trustee is going to assess the current situation, namely: i) the benefits of betraying; and ii) his relationship with the potential victim of betrayal. It is worth to note that Elangovan and Shapiro (1998)'s model has a third situational assessment component, the assessment of principles. In our model, we decided to incorporate this component into the betray propensity parameter described above; i.e., the principles of the trustee (related with his integrity) are used as triggers to the assessment of the situation, and not (directly) as one dimension of this assessment.

Assessing the Value of Betraying. The trustee assesses the benefits of betraying by taking into account both the utility associated with the new opportunity and the existence of a relevant sanction associated with the potential contract to betray. This sanction is considered *irrelevant* to the trustee if its value is smaller than a given (adjustable) percentage γ of the utility associated with the new opportunity. In this case, the value of betrayal is high. In order to reduce the complexity of the model, we chose three qualitative values for the value of betraying, as illustrated in equations 6.7-6.9.

$$Utility(p,x) - Sanction(c,x) - Utility(c,x) < \gamma_1/(1-\delta) \Rightarrow VBetrayal(c,x,low) \tag{6.7}$$

$$\begin{aligned} \gamma_1/(1-\delta) < Utility(p,x) - Sanction(c,x) - Utility(c,x) < \gamma_2/(1-\delta) \\ \Rightarrow VBetrayal(c,x,medium) \end{aligned} \tag{6.8}$$

$$\begin{aligned} \gamma_2/(1-\delta) < Utility(p,x) - Sanction(c,x) - Utility(c,x) \\ \Rightarrow VBetrayal(c,x,high) \end{aligned} \tag{6.9}$$

Assessing the Value of the Relationship. The trustee assesses the relationship with the potential victim by considering: (i) if the number of past contracts between both partners in the last σ units of time exceeds a minimum value λ (perspective of continuing the relationship, cf. Equation 6.10 where t denotes the current time step); and (ii) the existence of at least ξ other contracts in which the trustee is currently engaged (cf. Equation 6.11 where t denotes the current time step). The perceived value of the relationship is given in Table 6.5.

$$\sum_{i=1}^{\sigma} Contract(\_,x,y)^{t-i} > \lambda \Rightarrow PerspContinuity(x,y,high)\ . \tag{6.10}$$

$$\sum Contract(\_,x,\_)^{t} > \xi \Rightarrow HasOtherContracts(x)\ . \tag{6.11}$$

The decision to betray a partner or instead to keep his trust takes into consideration the assessment made by the trustee concerning the values of betrayal and relationship. In case there is more than one contract deemed to be betrayed, the trustee will only betray the one with less utility, provided that its allocated resources are enough to take into account the new contract.

Table 6.5: Value of the relationship.

PerspContinuity	HasOtherContracts	
	yes	*no*
high	medium	high
∄high	low	medium

Table 6.6: Trustee populations.

Population	ability	integrity
Heterogeneous	uniform distribution within [0.5, 1]	uniform distribution within $\{\delta_{low}, \delta_{medium}, \delta_{high}\}$
Low Integrity		δ_{low}
High Integrity		δ_{high}

In no contract is deemed to be betrayed, the trustee may still accept the new contract provided that enough free resources are available.

It is important to note that even new contracts may be later decided upon to be betrayed if another opportunity arises. Contracts are violated at enactment time, which means that the decision to betray is made much earlier than the *act* of betraying. Finally, while the trustee may decide not to betray a partner, he may still fail the contract if his ability is not good enough.

Experimental Setup

We have run three different set of experiments, each one using a different population of trustees, characterized by distinct values for the ability and integrity of the agents. Table 6.6 summarizes the different populations of trustees. In all experiments, we made $Tw_{x,y} = Sinalpha(E_{x,y})$. Also, the general configuration parameters of these experiments is presented in Table 6.7. The effective betrayal of contracts was configured probabilistically (see Table 6.8) taking into consideration the assessed values of the benefits of betraying and of the relationship.

Configuration of Trusters. The sanction value was calculated as $S = Ct - T/\zeta$. This formula provides the relationship between the trustworthiness of a trustee and the level of sanctions S that a truster will propose to be included in the contract. We start from the formulation of $Ct = T + S$, where for the reason explained before we reduced the weight of the trust parcel. Every truster had a value $Ra \in [0, 1]$ picked randomly at setup, and

Table 6.7: Configuration parameters (use of trust and sanctions).

δ_{low}, δ_{medium}, δ_{high}	0.1, 0.5, 0.9
$\rho_{\delta_{low}}$, $\rho_{\delta_{medium}}$, $\rho\delta_{high}$	0.5, 0.3, 0.1
γ_1, γ_2	0.0, 0.2
σ, λ, ξ, ζ	3, 2, 1, 4
# client agents	80
# provider agents	120
# rounds	80
# experiment repetitions	30
Sinalpha ω, λ_F and λ_V	$\pi/12$, 1.0 and -2.0
Sinalpha $lf(F)$ and $lf(V)$	0 and 1

Table 6.8: Betrayal probabilities.

	ValueBetrayal		
ValueRelationship	High	Medium	Low
High	0.5	0.0	0.0
Medium	1.0	0.0	0.0
Low	1.0	0.5	0.0

a value Pv also picked up randomly from a range of fixed minimum and maximum values. Tv is a dynamic value proposed by a trustee resulting from a specific contract negotiation.

Whenever betrayed, the truster resents the betrayal by ignoring any information regarding his previous activity, which has the effect of dropping his trustworthiness value to 0 (as assessed by the betrayed truster).

The final desideratum of the experimental component of our work was to evaluate the performance of the different combinations of trust and sanctions when applied to processes of selection of partners. Therefore, we defined the following types of trusters:

- None (`N`): The truster does not use sanctions nor trust.
- Sanctions (`S`): The truster uses sanctions but does not select partners based on trust.
- Trust (`T`): The truster uses trust to select partners but does not use sanctions.
- Trust and Sanctions (`T&S`): The truster uses trust both to select partners and to compute sanctions.

The population of trusters used in all sets of experiments followed a uniform distribution over the possible types described above.

Configuration of Trustees. In order to emulate the existence of a potential new opportunity (cf. Equation 6.6), all providers had a limited stock within a simulation round. The utility of a contract for a trustee was calculated by multiplying the dimension of the proposal (i.e. the quantity in the contract over the stock of the trustee) by the relevance of the price in the proposal.

Evaluation Variables. In these experiments, we observed the following evaluation variables:

- Δ_{sup}: number of different suppliers selected by all buyers in one round.
- $o+$: number of contracts with positive outcome (that were not violated or betrayed) in a round.
- O: number of opportunities to betray faced by the trustees. A trustee that has n active contracts when a new opportunity arises is confronted with n opportunities to betrayal.
- β: number of effective betrayals suffered by all trusters in one round.
- β/O : ratio of the number of effective betrayals to the number of opportunities to betray, indicating the effectiveness of the selection models in dissuading the trustees from betraying after identifying a new opportunity to betray.
- Ξ: number of contracts that were violated due to (lack of) ability. This variable is derived from the values of $o+$ and β.
- Σ: average sanction applied by all trusters to the contracts they establish in each round.

All variables took values in $[0, 1]$, all averaged over all rounds and all runs of the experiments.

Results for the Heterogeneous Population

The experimental results for the Heterogeneous population are shown in Table 6.9. The results in the Δ_{sup} evaluation metric show that agents that use trust in the selection process were less exploratory ($T : 0.880, T\&S :$

Table 6.9: Experimental results for Heterogeneous trustee population.

	Δ_{sup}	o^+	O	β	β/O	Ξ	Σ
N	0.963	0.708	0.127	0.028	0.217	0.265	0.000
S	0.966	0.721	0.126	0.017	0.132	0.262	0.230
T	0.880	0.833	0.307	0.055	0.180	0.112	0.000
T&S	0.878	0.846	0.309	0.044	0.143	0.180	0.096

0.878) than agents that did not use trust ($N : 0.963, S : 0.966$). In the same way, these trust-based agents were significantly more exposed to betray, as shown by variable O ($T : 0.307, T\&S : 309$), than the other agents ($N : 0.127, S : 0.126$).

One important variable to look at is the number of positive outcomes ($o+$), i.e., the number of contracts that were neither betrayed nor violated due to ability issues. We can observe in Table 6.9 that agents of type $T\&S$ outperformed the remaining agents in this variable. Indeed, agents of type $T\&S$ got more positive outcomes than T agents ($T : 0.833, T\&S : 0.846, t[1] = -5.78, p < 0.001$) and significantly better performance than agents of types $S(0.721)$ and $N(0.708)$.

Another important variable is the rate of materialized betrayals (β/O). Both of the truster types using sanctions achieved better performances on this issue (the results for these types are not statistically significantly different: $S : 0.132, S\&T : 0.143, t[1] = -0.85, p = 0.20$). Agents of type N performed worse than T agents ($N : 0.217, T : 0.180, t[1] = 3.18, p = 0.002$). Taking into consideration the two best truster types in the β/O metric, we can see that $T\&S$ agents used a much lighter sanction value Σ (0.096) that S agents did (0.230).

In terms of effective betrayals (β), S agents performed better than all other agents: agents of type N got more betrayals than agents of type S ($N : 0.028, S : 0.017, t = 5.51, p < 0.001$), and T agents got more betrayals (0.055) than $T\&S$ agents (0.044).

Finally, it is evident that trust-based models are more efficient in preventing violations due to lack of ability (Ξ) of trustees than the remaining models ($N : 0.265, S : 0.262, T : 0.112, T\&S : 0.180$). As can be observed, S agents could not do better than N agents concerning this issue.

Results for the Low Integrity Population

The experimental results for the Low Integrity populations are shown in Table 6.10. Similarly to what happened with the heterogeneous population, we verified that trust-based models were less explorative (Δ_{sup}) than the

Table 6.10: Experimental results for the Low Integrity population.

	Δ_{sup}	o^+	O	β	β/O	Ξ	Σ
N	0.962	0.694	0.203	0.052	0.256	0.254	0.000
S	0.961	0.709	0.207	0.034	0.167	0.257	0.219
T	0.851	0.783	0.555	0.116	0.210	0.102	0.000
T&S	0.861	0.808	0.562	0.089	0.159	0.104	0.093

remaining models ($N : 0.962, S : 0.961, T : 0.851, T\&S : 0.861$), which is inversely correlated with the opportunities to betray ($N : 0.203, S : 0.207, T : 0.555, T\&S : 0.562$). In the same way, we verified once again that the models that use sanctions were more effective regarding the rate of materialized betrayals (β/O), where the mean value obtained by agents of type S (0.167) was not significantly different from the equivalent mean value for $T\&S$ agents (0.159)($t[1] = 0.99, p = 0.17$). Concerning these truster types, $T\&S$ agents were able to use a lighter value of sanction Σ(0.093) than S agents (0.219). Agents that use trust but not sanctions could, even though, achieve a better value for β/O than agents that did not use trust neither sanctions ($N : 0.256, T : 0.210$).

In terms of the total value of positive outcomes ($o+$), both trust-based truster types outperformed the types that do not use trust in the selection process. $T\&S$ got more positive outcomes than T agents, benefiting from using sanctions ($T : 0.783, T\&S : 0.808, t[1] = -4.85, p < 0.001$). Concerning the trusters that did not use trust, S agents got a light advantage of about 2% over N agents ($N : 0.694, S : 0.709, t[1] = -4.04, p < 0.001$). Once again, it is evident that trust-based models are more efficient in preventing violations due to the lack of ability (Ξ) of trustees than the remaining models ($N : 0.254, S : 0.257, T : 0.102, T\&S : 0.104$).

The results of T agents and $T\&S$ agents were not statistically significantly different ($T : 0.102, T\&S : 0.104, t[1] = -1.10, p = 0.14$), and neither were the results of N and S agents ($N : 0.254, S : 0.257, t[1] = -0.85, p = 0.20$). Finally, S agents were the ones that got less betrays, in absolute terms ($N : 0.052, S : 0.034, T : 0.116, T\&S : 0.089$).

Results for the High Integrity Population

The experimental results for the High Integrity population are shown in Table 6.11. The results confirm the expected lower values of exploration of agents that use trust-based models when compared with the those that do not use trust to select partners ($N : 0.967, S : 0.967, T : 0.894, T\&S : 0.896$). T and $T\&S$ agents' exploration rate is about $92\% - 93\%$ of the rate of N

Table 6.11: Experimental results for the High Integrity population.

	Δ_{sup}	o^+	O	β	β/O	Ξ	Σ
N	0.967	0.732	0.046	0.006	0.120	0.263	0.000
S	0.967	0.735	0.045	0.002	0.051	0.262	0.217
T	0.894	0.866	0.108	0.011	0.107	0.122	0.000
T&S	0.896	0.870	0.109	0.010	0.092	0.120	0.086

and S agents, a value that is greater to what happened when using the population Low Integrity ($88\% - 90\%$). This may be explained by the fact that the population in general is more trustworthy (as the integrity of agents is higher and the ability remained unchanged) and therefore trust-based agents are less parochial in their selection choices.

The values of positive outcomes obtained using trust-based models were $18 - 19\%$ higher than those obtained by trusters that did not use trust. The mean values of $o+$ obtained by both T and $T\&S$ agents were not significantly different ($T : 0.866, T\&S : 0.870, t[1] = -1.55, p = 0.067$). This may indicate that when the opportunities to betrayal are low, the use of sanctions may not be relevant. The mean values of $o+$ obtained by N agents and S agents were not significantly different ($N : 0.732, S : 0.735, t[1] = -1.30, p = 0.10$).

As the population shows high integrity, the opportunities to betray are rather low for all types of agents ($N : 0.046, S : 0.045, T : 0.108, T\&S : 0.109$), which translates also in low values of effective betrayals. We verified that N agents suffered more betrayals than S agents ($N : 0.006, S : 0.002, t[1] = 2.76, p = 0.005$), and that the mean values of betrayals obtained by both T and $T\&S$ agents were not significantly different ($T : 0.011, S\&T : 0.010, t[1] = 2.11, p = 0.021$). In any case, all these values are almost residual, even for trust-based models.

Once again, we verified that the trusters that use sanctions were more effective regarding the rate of materialized betrayals (β/O), where the mean value obtained by S agents was not significantly different from the equivalent mean value for $T\&S$ agents ($S : 0.051, S\&T : 0.092, t[1] = -2.20, p = 0.018$). $T\&S$ agents were able to use a lighter value of sanction (0.086) than S agents (0.217). Finally, it is evident that trust-based models were more efficient in preventing violations due to the lack of ability (Xi) of trustees than the remaining models ($N : 0.263, S : 0.262, T : 0.122, S\&T : 0.120$).

Discussion

The results presented in the previous section for the different populations have shown that $T\&S$ is the best selection model concerning the total num-

ber of positive outcomes, i.e., the number of contracts that were neither failed due to lack of ability of trustees nor betrayed by them. This is due to the combined effect of using trust - which proved to be very effective in avoiding trustees with lower ability - and sanctions - which have shown to have an important role in persuading the trustees to maintain the status quo after identifying an opportunity to betray.

In our study, we settle an extremely complex scenario, with different models of behavior for trusters and trustees, where both models were inspired in theoretical works on normative control and trust. However, the novelty of such an approach came with a price: the resulting model had a great number of variables, potentially influencing each other, which hardened the analysis of the experimental results. For instance, a not irrelevant bias of our model is related to the fact that trusters that selected partners based on utility and trust were more exposed to potential betrayals (as shown by the opportunities of betrayal variable) than trusters that selected based only on the utility of the proposals. Indeed, the fact that the latter explored more partners implied that each trustee had fewer active contracts at one time, and consequently less potential contracts to consider betraying when a new opportunity arose.

The mentioned bias reflected directly on the results of the betray variable. Therefore, it is probably wiser to take into consideration the results of the β/O variable (i.e., how many opportunities of betrayal do materialize into an effective betrayal) rather than the results on the betrayal variable, in order to understand the effectiveness of each selection model in preventing betrayals. We verified that both selection models that use sanctions (S and $T\&S$) have shown similar performance in this variable. However, the $T\&S$ model had the additional advantage of using lighter sanctions. This happened, once again, due to the complementary action of trust and sanctions: by selecting the most trustworthy agents and considering that sanctions were drafted (also) taking into account the perceived trustworthiness of trustees, the value of the applied sanction was reduced.

The study we presented in this section is novel, in the sense that it experimentally analyzed the combined effect of trust and normative control (in the form of sanctions) in the process of selecting partners. Also, it used models of behavior more realistic than the models generally used to test trust and norms (standalone), which are generally probabilistic and static. However, our model and our work present limitations that must be taken into consideration in future work. For instance, even though the behavioral models drink from theoretical insights on trust and sanctions, the overall model is not empirically grounded.

A second limitation concerns the simplifications that were done to (Elangovan and Shapiro, 1998)'s model in order to reduce its complexity. A way to more closely follow the model without introducing unbearable experimental complexity must be addressed in the future.

Based on the conditions of these experiments, we were motivated to continue this work in order to fully accept Hypothesis 4.

6.4 Concluding Remarks

In this chapter, we presented the ANTE platform, an agent-based framework for the negotiation of agreements that is being developed at the LIACC group. Taking this platform as basis, we identified three key stages where trust can be used to improve the performance of the negotiation process: i) in the selection of partners; ii) in the automatic drafting of contracts; and iii) in the generation of contractual evidences.

Also in the context of the ANTE platform, we performed two different studies about the use of trust as an agreement technology. The first evaluated the benefits that exist when using computational trust to enhance the outcome of automatic negotiations. The second study was conducted in order to test if the use of trust in contracting agent systems could decrease the weight of sanctions without jeopardizing the efficiency of normative control in promoting the compliance of agents. The results are still preliminary but promissory.

Chapter 7

Conclusions and Future Work

In this chapter, we outline the main achievements of this thesis and the questions that remain unanswered. We start by presenting an overview of the work that we have developed in the scope of this thesis, in Section 7.1. Then, we remember and summarize the main contributions of this thesis, in Section 7.2. Finally, Section 7.3 presents the limitations of our approach, the final conclusions and future work.

7.1 Thesis Summary

We started the work of this thesis motivated by the need of developing a model of computational trust for the ANTE platform that would be more sophisticated than the one already existing at the LIACC Laboratory. After some months of research studying the computational trust approaches existing at that time, we realized that most of them were focused on developing new trustworthiness estimators using different algorithms for the aggregation of trust-based evidence. Other branch of research with strong development at this time was the management of reputation information and its inclusion in computational trust models. However, we soon felt that a large number of the proposals on computational trust were not grounded on theoretical aspects of trust and seemed to ignore the vast multi-disciplinary literature on trust and associated concepts.

Our exploratory study on theoretical aspects of trust provided us invaluable insights that we would not have acquired if our review of literature had focused only on computational models of trust and reputation. We mention the most important of these insights and their effect on the contributions of

our work next:

- *Trust and trustworthiness are distinct constructs*, although sometimes they tend to be used interchangeably in the computational trust and reputation literature. We make a clear distinction of both constructs in the situation-based, social-based trust framework, SOLUM, presented in Chapter 4.

- *There are other antecedents to trust than trustworthiness*, such as the truster's disposition and emotional state. Once again, this is generally ignored in computational trust research. The exceptions are the works by (Kelton et al., 2008; Adali et al., 2011) that refer the truster's disposition in their conceptual models. We consider both antecedents in the SOLUM framework.

- *Trustworthiness is a multi-dimensional construct*, and its different dimensions can be roughly grouped into ability, integrity and benevolence. However, most of the existent computational trust approaches ignore this fact. The socio-cognitive model of trust of Castelfranchi and Falcone (2010) was probably the first of these approaches to consider the multi-dimensionality of trust and to implement it. However, it considers the existence of specific information about each one of these dimensions. In our work, we consider the existence of these dimensions in the SOLUM framework, and present *Integrity Tuner* and *Social Tuner*, two computational components that respectively estimate the integrity and benevolence of trustees from the available evidence on these trustees.

- *Trust is, at least, a quaternary relation*: the trusting judgment concerns a truster, a trustee, a task, and the contextual situation inherent to this judgment. Although a significant number of computational trust models that we evaluated ignore the situation-awareness of trust, there are indeed situation-aware approaches to computational trust (e.g. Rehak et al., 2006; Tavakolifard et al., 2008). However, to the best of our knowledge, these approaches are based on the definition of similarity distances between situations, which means that these situations and distances must be known and predefined *apriori*. To avoid this constraint, we developed *Contextual Fitness*, a simple algorithm based on the information gain metric that extracts tendencies of past behavior instead of defining all possible situations, and that can be applied to any situation-less computational model.

- *Trust is a social concept.* This means that the stage of the relationship between truster and trustee is of vital importance when judging trust, and that understanding the stage of the relationship is paramount to combine the contribution of the different trustworthiness dimensions when estimating the agents' trustworthiness. To the best of our knowledge, we present in this thesis the first and unique approach to computational trust that considers the current stage of the relationship when making trust judgments. Not only we define ways to estimate the current level of the trustees' integrity and benevolence, but we also present two preliminary approaches to combine these dimensions in different stages of the truster-trustee relationship, in Section 4.8.

All these insights were reflected or have influenced the six research questions that we described in Section 1.2.1, and therefore they guided all research of this thesis. The next part of our work consisted in studying the best way to address the research questions we formulated and to transpose the insights just described into a computational model of trust. We searched for practical and simple solutions that could be applied to real problems and applications, avoiding the complexity of richer solutions that would be unfeasible to implement with the current knowledge and technology. With this in mind, we opted to implement some desirable features we had identified and opted to postpone others to future work. For instance, taking into consideration the time constrains of the thesis, we had to leave aside the consideration of the truster's disposition and emotional state (including the emotions felt after a betrayal) in the trust function.

Taking again into consideration the problem-oriented nature of our research, we made another option that guided our research throughout this thesis: we assumed that, in artificial societies, the detailed information about the ability, integrity and benevolence of specific trustees would not be easily available to trusters. Then, we opted to build our trust-based components in a way that they would be able to extract this information from the available, structured set of evidence. This clearly distinguishes our work from the socio-cognitive model of trust of Castelfranchi and Falcone (2010), which, to the best of our knowledge, is the only computational trust model that does *implement* the different antecedents of trustworthiness. In fact, this model assumes that specific and detailed information on these dimensions exists. With this, we are not saying that this type of information should not be considered, but that it is probable that in open and dynamic agent-based systems this information should not be abundant. Then, in these circumstances, trusters could not do any better than to infer the information about

the ability, integrity and benevolence of agents from the unspecific, structured set of evidence, as we propose, to complement their knowledge base.

In the remaining work of this thesis, we instantiated the SOLUM framework and developed our computational components: *Sinalpha*, *Contextual Fitness*, *Integrity Tuner* and *Social Tuner*. We also developed two different approaches to combine the outcomes of these components into trustworthiness estimations. These components allowed us to evaluate the four hypotheses that we formulated in Chapter 4, which were related with the six research questions enumerated in Chapter 1. We address the evaluation of these hypotheses next.

First Hypothesis

The first hypothesis we have formulated stated that the extraction of benevolence based information from the set of evidence on the trustee under evaluation and its use in adequate stages of the relationship between truster and trustee improves the reliability of the estimation of this trustee's trustworthiness.

From the experimental evaluation of the *Social Tuner* component, we were able to conclude that trusters that used this component were able to make more reliable trust decisions than trusters that did not use this component, in the conditions of the experiments; that the benefits of *Social Tuner* were equally evident when this component was used in conjunction with different other types of trustworthiness estimators besides *Sinalpha*, namely, the ones presented in (Jonker and Treur, 1999; Jøsang and Ismail, 2002); and that trusters that used the *Social Tuner* component in environments not driven by benevolence-based behaviors did not made worse trust-based decisions than those not using this component. We have also evaluated the use of *Social Tuner* when combined with other components, assigning different weights to the use of the benevolence-based information taking into consideration the stage of the relationship between truster and trustee. The results of this evaluation, even preliminary, showed that trusters that implement such strategy got better results than trusters that did not use it.

From all results, we were able to confirm the trustfulness of this first hypothesis, in the conditions of the experiments.

Second Hypothesis

The second hypothesis stated that the use of proper computational techniques that enable to extract contextual information from the set of evidence on the trustee under evaluation improves the reliability of the estimation of

this trustee's trustworthiness. The consequent reliability of the trust decision is improved even when the available evidence is scarce.

From the experimental evaluation of the *Contextual Fitness* component, we were able to conclude that the use of these techniques that enable to extract contextual information from the set of available evidence in fact improved the reliability of the estimation of this trustee's trustworthiness; that the consequent reliability of the trust decision was improved even when the available evidence was scarce; that *Contextual Fitness* performed better than an alternative approach for computational situation-aware trust based on pre-defined measures of similarity between situations; that *Contextual Fitness* supported the search for more desirable partners in mutualistic terms without jeopardizing the trust-based selection decisions; and that the benefits of *Contextual Fitness* were equally evident when this component was used in conjunction with different other types of trustworthiness estimators besides *Sinalpha*, namely, the well known proposals of Jøsang and Ismail (2002) and Jonker and Treur (1999).

From all results, we were able to confirm the trustfulness of this second hypothesis, in the conditions of the experiments.

Third Hypothesis

The third hypothesis stated that the extraction of integrity-based information from the set of evidence on the trustee under evaluation improves the reliability of the estimation of this trustee's trustworthiness.

From the experimental evaluation of the *Integrity Tuner* component, we were able to conclude that trusters that added this component to the trust-based evidence aggregator were able to make more reliable trust decisions than the ones that did not use this component. This was valid for all three evidence aggregators that we considered in the experiments. In the same way, the improved reliability in trustworthiness estimation was observed even when the populations of agents used in the experiments did not present specific characteristics concerning their dispositional integrity.

From all results, we were able to confirm the trustfulness of this third hypothesis, in the conditions of the experiments.

Fourth Hypothesis

The fourth hypothesis stated that the use of trust in contracting agent systems decreases the weight of sanctions without jeopardizing the efficiency of normative control in promoting the compliance of agents.

This hypothesis was related with a complementary branch of research of this thesis, conducted in the scope of project PTDC/EIA-EIA/104420/2008 sponsored by FCT. To evaluate this hypothesis, we settle an experimental testbed that included the accommodation of the SOLUM computational trust model into the ANTE platform, as an autonomous service of this framework, and its integration with two other fundamental services of the framework: the negotiation facilitator and the normative environment.

The results we obtained are still preliminary, not allowing to fully confirm the truthfulness of the hypothesis. However, these results were promissory. More interesting, our work allowed to explore a new research branch on the interplay between computational trust and computational normative environments, and we believe that our study, as well as the obtained results, may serve as baseline to others, more elaborated studies on this very important topic.

7.2 Research Contributions

The research contributions of this thesis have being referred to throughout this thesis, but can be summarized in the following items:

- The presentation of a set of propositions about the nature and essence of the trust construct, with a special emphasis to the social nature of trust. These propositions have guided our work throughout this thesis, and we hope they can also be useful to other researchers, specially those who are giving the first steps on the field of computational trust.
- The presentation of SOLUM, a generic framework for reasoning about social trust that can be instantiated into different models tailored to specific domains.
- The proposal of practical implementations of parts of this framework. Namely, the *Contextual Fitness*, *Social Tuner*, and *Integrity Tuner* have shown to improve the reliability of the trustworthiness estimation, either individually or as a group. In the same way, we offer an innovative thinking about the aggregation of information about ability, integrity and benevolence, taking into consideration the situation and the stage of the relationship existing between truster and trustee. With this, we advanced the state-of-the-art on computational trust.
- An introductory and innovative study on the use of trust in agent-based systems for contracting, which may serve as baseline to others, more elaborated studies on this very important topic.

- The presentation of a practical approach where computational trust is used as an agreement technology in a framework for automatic agreements that presents additional services of automatic negotiation and normative environment.

7.3 Limitations and Future Work

When addressing a topic as vast and recent as computational trust, several issues are guaranteed to be put aside and some assumptions and compromises need to be taken. In fact, all research involves the simplification of the problem in investigation. Our work is not an exception.

First, we were not able to prove the benefits of *Sinalpha*. Moreover, the often referred observance of this component's algorithm to the asymmetry and perseverance of trust would be more wisely applicable to the trust function $Tr_{x,y}$ of the SOLUM framework than to the estimation of the global competence of the trustee. In the same way, the implementation of the ability evaluation function $ab_{x,y}$ using a common trustworthiness estimator needs to be reevaluated, and the ability dimension of trustworthiness needs to be better grounded.

Second, we were able to confirm the truthfulness of three of the four formulated hypotheses, in the conditions of the experiments we performed. Although we believe that the obtained results might be generalizable to other conditions, further experiments and evaluation methods are needed to fully confirm the hypotheses. In a related topic, our model of agents' behavior is not empirical. Although we believe that by allowing the creation of evolving structures of agents relationships, our model is more realistic and consequently more useful to the evaluation of computational trust approaches than models where agents fulfill or violate their agreements based on some simple probability, the validity of the results obtained with this model are restricted to the conditions of the model.

Third, some of the work related to the trust-based computational components we presented is based on heuristics, and involved the definition of parameters whose values were prefixed based on these heuristics. We believe that these components shall be further developed in order to allow for some of these parameter values to be automatically learned.

Fourth, as we have already mentioned, we intend to instantiate additional components of the SOLUM framework, namely, the aggregation of reputation-like information and of specific and precise information about a given trustworthiness dimension of the trustee under evaluation. One possibility for such aggregation would be to try the node-based, FCM-based

approach proposed by (Castelfranchi and Falcone, 2010).

Fifth, concerning the integration of SOLUM in the ANTE platform, we intend to evolve the computational trust service and its application interface, in order to allow it to be decentralized. In the some way, we would like to explore the relation between the consequences of a betrayal and the emotional state of the truster, which feeds evaluation function $Tr_{x,y}$.

Finally, at a computational efficiency point of view, we intend to study methods for condensing the past evidence, using, for instance, the concept of epochs presented in (Ruohomaa and Kutvonen, 2008). In the same way, proved the concept, we intend to explore online classification algorithms that would allow to a more efficient implementation of *Contextual Fitness*.

Bibliography

A. Abdul-Rahman and S. Hailes. Supporting Trust in Virtual Communities. In *Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 6 - Volume 6*, HICSS '00, pages 6007–, Washington, DC, USA, 2000. IEEE Computer Society. ISBN 0-7695-0493-0.

G. Abowd, A. Dey, P. Brown, N. Davies, M. Smith, and P. Steggles. Towards a better understanding of context and context-awareness. In H.-W. Gellersen, editor, *Handheld Ubiquit. Comput.*, volume 1707 of *LNCS*, pages 304–307. Springer Berlin, 1999. ISBN 978-3-540-66550-2.

S. Adali, W. A.Wallace, Y. Qian, P. Vijayakumar, and M. P. Singh. A unified framework for trust in composite networks. In *Proceedings of the 14th AAMAS Workshop on Trust in Agent Societies*, pages 1–12, May 2011. Taipei.

P. D. Allison. The cultural evolution of beneficent norms. *Social Forces*, 71 (2):279–301, 1992.

P. Alves, P. Campos, and E. Oliveira. Modeling the trustworthiness of a supplier agent in a b2b relationship. In L. Camarinha-Matos, L. Xu, and H. Afsarmanesh, editors, *Collaborative Networks in the Internet of Services*, volume 380 of *IFIP Advances in Information and Communication Technology*, pages 675–686. Springer Boston, 2012. ISBN 978-3-642-32774-2.

R. Bachmann. Trust, power and control in trans-organizational relations. *Organization Studies*, 22(2):341–369, 2001.

B. Banerjee, A. Biswas, M. Mundhe, S. Debnath, and S. Sen. Using bayesian networks to model agent relationships. *Applied Artificial Intelligence*, 14 (9):867–879, 2000.

T. E. Becker. Development and validation of a situational judgment test of

employee integrity. *International Journal of Selection and Assessment*, 13 (3):225–232, 2005. ISSN 1468-2389.

F. L. Bellifemine, G. Caire, and D. Greenwood. *Developing Multi-Agent Systems with JADE*. Wiley Series in Agent Technology. John Wiley & Sons, Ltd, 2007.

R. Bhattacharya, T. M. Devinney, and M. M. Pillutla. A formal model of trust based on outcomes. *The Academy of Management Review*, 23(3): 459–472, Jul. 1998.

G. Boella, J. Hulstijn, Y.-H. Tan, and L. van der Torre. Transaction trust in normative multiagent systems. In J. Sabater, editor, *8th Workshop on Trust, Privacy, Deception and Fraud in Agent Societies (Trust'05)*, Utrecht, The Netherlands, 2005.

G. Boella, L. van der Torre, and H. Verhagen. Introduction to the special issue on normative multiagent systems. *Autonomous Agents and Multi-Agent Systems*, 17(1):1–10, 2008.

T. Bosse, C. M. Jonker, J. Treur, and D. Tykhonov. Formal analysis of trust dynamics in human and software agent experiments. In *Proc. CIA '07*, pages 343–359, Berlin/Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-75118-2.

T. J. Bouchard and M. McGue. Genetic and environmental influences on human psychological differences. *J. Neurobiology*, 54(1):4–45, 2003. ISSN 1097-4695.

C. Burnett. *Trust Assessment and Decision-Making in Dynamic Multi-Agent Systems*. PhD thesis, University of Aberdeen, 2011.

C. Burnett, T. J. Norman, and K. Sycara. Bootstrapping trust evaluations through stereotypes. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1 - Volume 1*, AAMAS '10, pages 241–248, Richland, SC, 2010. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-0-9826571-1-9.

D. Bzdok, R. Langner, S. Caspers, F. Kurth, U. Habel, K. Zilles, A. Laird, and S. Eickhoff. Ale meta-analysis on facial judgments of trustworthiness and attractiveness. *Brain Structure and Function*, 215:209–223, 2011. ISSN 1863-2653.

L. Camarinha-Matos, H. Afsarmanesh, H. Löh, F. Sturm, and M. Ollus. A strategic roadmap for advanced virtual organizations. In L. Camarinha-Matos and H. Afsarmanesh, editors, *Collaborative Networked Organizations*, pages 289–312. Springer US, 2004. ISBN 978-1-4020-7833-0.

L. Camarinha-Matos, I. Silveri, H. Afsarmanesh, and A. Oliveira. Towards a framework for creation of dynamic virtual organizations. In L. Camarinha-Matos, H. Afsarmanesh, and A. Ortiz, editors, *Collaborative Networks and Their Breeding Environments*, volume 186 of *IFIP International Federation for Information Processing*, pages 69–80. Springer Boston, 2005. ISBN 978-0-387-28259-6.

J. Carbo, J. Molina, and J. Davila. Trust management through fuzzy reputation. *International Journal of Cooperative Information Systems*, 12(1): 135 – 55, 2003. ISSN 0218-8430.

K. Carley, D. Park, and M. Prietula. Agent honesty, cooperation and benevolence in an artificial organization. Technical report, Institute for Software Research, 1993.

C. Castelfranchi and R. Falcone. Principles of trust for mas: Cognitive anatomy, social importance, and quantification. In *Proceedings of the 3rd International Conference on Multi Agent Systems*, ICMAS '98, pages 72–79, Washington, DC, USA, 1998. IEEE Computer Society. ISBN 0-8186-8500-X.

C. Castelfranchi and R. Falcone. *Trust Theory: A Socio-Cognitive and Computational Model*. Wiley Series in Agent Technology. John Wiley & Sons Ltd., Chichester, 2010.

C. Castelfranchi, R. Falcone, and G. Pezzulo. Trust in information sources as a source for trust: A fuzzy approach. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '03, pages 89–96, New York, NY, USA, 2003. ACM. ISBN 1-58113-683-8.

B. Christianson and W. S. Harbison. Why isn't trust transitive? In *Proceedings of the International Workshop on Security Protocols*, pages 171–176, London, UK, UK, 1997. Springer-Verlag. ISBN 3-540-62494-5.

J. A. Colquitt, B. A. Scott, and J. A. LePine. Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, 92:909–927, July 2007.

B. S. Connelly, S. O. Lilienfeld, and K. M. Schmeelk. Integrity tests and morality: Associations with ego development, moral reasoning, and psychopathic personality. *International Journal of Selection and Assessment*, 14(1):82–86, March 2006.

R. Conte and M. Paolucci. *Reputation in Artificial Societies: Social Beliefs for Social Order*. Kluwer Academic Publishers, Norwell, MA, USA, 2002. ISBN 1402071868.

D. Cox, M. La Caze, and M. Levine. Integrity. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2012 edition, 2012.

G. Cvetkovich, M. Siegrist, R. Murray, and S. Tragesser. New information and social trust: Asymmetry and perseverance of attributions about hazard managers. *Risk Analysis*, 22(2):359–367, 2002. ISSN 1539-6924.

A. Danek, J. Urbano, A. P. Rocha, and E. Oliveira. Engaging the dynamics of trust in computational trust and reputation systems. In *Proceedings of the 4th KES international conference on Agent and multi-agent systems: technologies and applications, Part I*, KES-AMSTA'10, pages 22–31, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-13479-3, 978-3-642-13479-1.

V. D. Dang. *Coalition Formation and Operation in Virtual Organisations*. PhD thesis, University of Southampton, 2004.

T. K. Das and B. Teng. Between trust and control: Developing confidence in partner cooperation in alliances. *Academy of Management Review*, 23 (3):491–512, 1998.

P. Dasgupta. Trust as a commodity. In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, pages 49–72. Department of Sociology, University of Oxford, 2000.

R. Demolombe. Transitivity and propagation of trust in information sources: an analysis in modal logic. In *Proceedings of the 12th international conference on Computational logic in multi-agent systems*, CLIMA'11, pages 13–28, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-22358-7.

F. Dignum and C. Sierra, editors. *Agent Mediated Electronic Commerce, The European AgentLink Perspective.*, London, UK, 2001. Springer-Verlag. ISBN 3-540-41671-4.

T. Dimitrakos. System models, e-risks and e-trust. In B. Schmid, K. Stanoevska-Slabeva, and V. Tschammer, editors, *Towards the E-Society*, volume 74 of *IFIP International Federation for Information Processing*, pages 45–58. Springer Boston, 2002. ISBN 978-0-7923-7529-6.

K. Dion, E. Berscheid, and E. Walster. What is beautiful is good. *Journal of Personality and Social Psychology*, 24(3):285–290, Dec 1972.

J. R. Dunn and M. E. Schweitzer. Feeling and believing: the influence of emotion on trust. *Journal of Personality and Social Psychology*, 88(5): 736–748, 2005.

A. R. Elangovan and D. L. Shapiro. Betrayal of trust in organizations. *The Academy of Management Review*, 23(3):547–566, July 1998.

G. Elofson. Developing trust with technology: An exploratory study. In *Proceedings of the first International Workshop on Trust*, pages 125–139, 1998.

E. Erriquez, W. van der Hoek, and M. Wooldridge. An abstract framework for reasoning about trust. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '11, pages 1085–1086, Richland, SC, 2011. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 0-9826571-7-X, 978-0-9826571-7-1.

M. Esteva, J.-A. Rodríguez-Aguilar, C. Sierra, P. Garcia, and J. Arcos. On the formal specification of electronic institutions. In F. Dignum and C. Sierra, editors, *Agent Mediated Electronic Commerce*, volume 1991 of *Lecture Notes in Computer Science*, pages 126–147. Springer Berlin / Heidelberg, 2001. ISBN 978-3-540-41671-5.

A. Fabregues and J. Madrenas-Ciurana. Srm: a tool for supplier performance. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '09, pages 1375–1376, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-0-9817381-7-8.

E. J. Finkel, C. E. Rusbult, M. Kumashiro, and P. A. Hannon. Dealing with betrayal in close relationships: Does commitment promote forgiveness? *Journal of Personality and Social Psychology*, 82(6):956–974, Jun 2002.

J. Fitness. Betrayal, rejection, revenge, and forgiveness: An interpersonal script approach. *Interpersonal rejection*, pages 73–103, 2001.

M. Foddy, M. J. Platow, and T. Yamagishi. Group-based trust in strangers. *Psychological Science*, 20(4):419–422, 2009.

R. H. Frank. *Passions within reason: The strategic role of the emotions.* W W Norton & Co., New York, NY, US, xiii edition, 1988. 304 pp.

D. Gambetta. Can we trust trust? In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations, electronic edition*, chapter 13, pages 213–237. Department of Sociology, University of Oxford, 2000.

J. Golbeck, B. Parsia, and J. Hendler. Trust networks on the semantic web. In M. Klusch, A. Omicini, S. Ossowski, and H. Laamanen, editors, *Cooperative Information Agents VII*, volume 2782 of *Lecture Notes in Computer Science*, pages 238–249. Springer Berlin / Heidelberg, 2003. ISBN 978-3-540-40798-0.

M. Grabowski and K. H. Roberts. Risk mitigation in virtual organizations. *Organization Science*, 10:704–721, June 1999. ISSN 1526-5455.

N. Griffiths. Task delegation using experience-based multi-dimensional trust. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, AAMAS '05, pages 489–496, New York, NY, USA, 2005. ISBN 1-59593-093-0.

D. Grossi, H. Aldewereld, and F. Dignum. Ubi lex, ibi poena: Designing norm enforcement in e-institutions. In P. Noriega, J. Vázquez-Salceda, G. Boella, O. Boissier, V. Dignum, N. Fornara, and E. Matson, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems II*, volume LNAI 4386, pages 101–114. Springer, 2007.

N. Gujral, D. DeAngelis, K. K. Fullam, and K. S. Barber. Modeling multi-dimensional trust. In *Procs. of The Workshop on Trust in Agent Societies at AAMAS-2006*, pages 35–41, May 2006.

J. Haller. A bayesian reputation system for virtual organizations. In H. Gimpel, N. R. Jennings, G. E. Kersten, A. Ockenfels, C. Weinhardt, W. Aalst, J. Mylopoulos, M. Rosemann, M. J. Shaw, and C. Szyperski, editors, *Negotiation, Auctions, and Market Engineering*, volume 2 of *Lecture Notes in Business Information Processing*, pages 171–178. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-77554-6.

R. Hardin. Trust and society. In G. Galeotty, P. Slamon, and R. Wintrobe, editors, *Competition and structure. The political economy of collective*

decisions: Essays in honor of Albert Breton, pages 17–46. Cambridge University Press, 2000.

R. Hardin. Conceptions and explanations of trust. In K. S. Cook, editor, *Trust in society*, volume 2, pages 3–39. Russell Sage foundation series on trust, New York, NY, US, 2001.

R. Hardin. *Trust and trustworthiness*. The Russell Sage Foundation series on trust. Russell Sage Foundation, New York, NY, US, 2002.

R. Hardin. Distrust: Manifestation and management. In R. Hardin, editor, *Distrust*, pages 3–33. Russell Sage Foundation, 2004.

C. Heimer. Solving the problem of trust. In K. S. Cook, editor, *Trust in Society*, pages 40–88. Russell Sage Foundation Series on Trust, 2001.

R. Hermoso, H. Billhardt, and S. Ossowski. Dynamic evolution of role taxonomies through multidimensional clustering in multiagent organizations. In J.-J. Yang, M. Yokoo, T. Ito, Z. Jin, and P. Scerri, editors, *Principles of Practice in Multi-Agent Systems*, volume 5925 of *Lecture Notes in Computer Science*, pages 587–594. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-11160-0.

A. Herzig, E. Lorini, J. F. Hübner, and L. Vercouter. A logic of trust and reputation. *Logic Journal of the IGPL*, 18(1):214–244, 2010.

A. Hirschman. Against parsimony: Three easy ways of complicating some categories of economic discourse. *Economics and Philosophy*, 1:7–21, 1985.

M. Hoogendoorn, S. W. Jaffry, and J. Treur. An adaptive agent model estimating human trust in information sources. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 02*, WI-IAT '09, pages 458–465, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3801-3.

M. Hoogendoorn, S. W. Jaffry, and J. Treur. Cognitive and neural modeling of dynamics of trust in competitive trustees. *Cognitive Systems Research*, 14(1):60–83, 2012. ISSN 1389-0417.

T. D. Huynh, N. R. Jennings, and N. R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13:119–154, September 2006. ISSN 1387-2532.

R. D. Ireland and J. W. Webb. A multi-theoretic perspective on trust and power in strategic supply chains. *Journal of Operations Management*, 25 (2):482 – 497, 2007. ISSN 0272–6963. Special Issue Evolution of the Field of Operations Management SI/ Special Issue Organisation Theory and Supply Chain Management.

G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 538–543, New York, NY, USA, 2002. ACM. ISBN 1-58113-567-X. doi: 10.1145/775047.775126.

C. M. Jonker and J. Treur. Formal analysis of models for the dynamics of trust based on experiences. In F. Garijo and M. Boman, editors, *Multi-Agent System Engineering*, volume 1647 of *Lecture Notes in Computer Science*, pages 221–231. Springer Berlin / Heidelberg, 1999. ISBN 978-3-540-66281-5.

A. Jøsang. A logic for uncertain probabilities. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 9(3):279–311, June 2001. ISSN 0218-4885.

A. Jøsang and R. Ismail. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, Bled, Slovenia, June 2002. 17-19 June 2002.

A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decis. Support Syst.*, 43:618–644, March 2007. ISSN 0167-9236.

J. J. Jung. Ontology-based context synchronization for ad hoc social collaborations. *Know.-Based Syst.*, 21(7):573–580, Oct. 2008. ISSN 0950-7051. doi: 10.1016/j.knosys.2008.03.015.

R. Jurca and B. Faltings. An incentive compatible reputation mechanism. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, AAMAS '03, pages 1026–1027, New York, NY, USA, 2003. ACM. ISBN 1-58113-683-8.

Y. Karabulut and J. Sairamesh. Market study - wp15 exploitation. Technical report, TrustCoM Deliverable 7, 2005. v.2.3.

K. Kelton, K. R. Fleischmann, and W. A. Wallace. Trust in digital information. *J. Am. Soc. Inf. Sci. Technol.*, 59(3):363–374, Feb. 2008. ISSN 1532-2882.

F. Kerschbaum, J. Haller, Y. Karabulut, and P. Robinson. Pathtrust: A trust-based reputation service for virtual organization formation. In K. Stølen, W. Winsborough, F. Martinelli, and F. Massacci, editors, *Trust Management*, volume 3986 of *Lecture Notes in Computer Science*, pages 193–205. Springer Berlin / Heidelberg, 2006.

T. Kiyonari, T. Yamagishi, K. S. Cook, and C. Cheshire. Does trust beget trustworthiness? trust and trustworthiness in two games and two cultures: A research note. *Social Psychology Quarterly*, 69(3):270–283, Sep. 2006.

S. König, S. Hudert, T. Eymann, and M. Paolucci. Towards reputation enhanced electronic negotiations for service oriented computing. In *IEEE Joint Conference on E-Commerce Technology (CEC'08) and Enterprise Computing, E-Commerce and E-Services (EEE'08)*, pages 285–290, Crystal City, Washington D.C., USA, 2008. IEEE Computer Society.

T. R. Koscik and D. Tranel. The human amygdala is necessary for developing and expressing normal interpersonal trust. *Neuropsychologia*, 49(4):602 – 611, 2011. ISSN 0028-3932.

R. V. Lapshin. Analytical model for the approximation of hysteresis loop and its application to the scanning tunneling microscope. *Review of Scientific Instruments*, 66(9):4718–4730, September 1995. ISSN 0034-6748.

D.-J. Lee, M. J. Sirgy, J. R. Brown, and M. M. Bird. Importers' benevolence toward their foreign export suppliers. *Journal of the Academy of Marketing Science*, 32(1):32–48, 2004.

D.-J. Lee, I. Jeong, H. T. Lee, and H. J. Sung. Developing a model of reciprocity in the importerŰexporter relationship: The relative efficacy of economic versus social factors. *Industrial Marketing Management*, 37(1): 9–22, 2008.

D. Z. Levin, R. Cross, L. C. Abrams, and E. L. Lesser. Trust and knowledge sharing: A critical combination. In E. Lesser and L. Prusak, editors, *Creating Value with Knowledge : Insights from the IBM Institutue for Business Value*, pages 36–43. Oxford University, Oxford, 2004.

D. Z. Levin, E. M. Whitener, and R. Cross. Perceived trustworthiness of knowledge sources: The moderating impact of relationship length. *Journal of Applied Psychology*, 91(5):1163–1171, September 2006.

J. D. Lewis and A. Weigert. Trust as a social reality. *Social Forces*, 63(4): 967–985, June 1985.

H. Lopes Cardoso. *Electronic Institutions with Normative Environments for Agent-based E-contracting.* PhD thesis, University of Porto, September 2010.

H. Lopes Cardoso and E. Oliveira. Virtual enterprise normative framework within electronic institutions. In M.-P. Gleizes, A. Omicini, and F. Zambonelli, editors, *Engineering Societies in the Agents World V*, volume 3451 of *Lecture Notes in Computer Science*, pages 898–898. Springer Berlin / Heidelberg, 2005. ISBN 978-3-540-27330-1.

H. Lopes Cardoso and E. Oliveira. Directed deadline obligations in agent-based business contracts. In J. Padget, A. Artikis, W. Vasconcelos, K. Stathis, V. Torres da Silva, E. Matson, and A. Polleres, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems V*, LNAI 6069, pages 225–240. Springer, 2010.

H. Lopes Cardoso and E. Oliveira. Social control in a normative framework: An adaptive deterrence approach. *Web Intelli. and Agent Sys.*, 9(4):363–375, 2011. ISSN 1570-1263.

H. Lopes Cardoso, J. Urbano, P. Brandão, A. Rocha, and E. Oliveira. Ante: Agreement negotiation in normative and trust-enabled environments. In Y. Demazeau, J. P. Müller, J. M. C. Rodríguez, and J. B. Pérez, editors, *Advances on Practical Applications of Agents and Multi-Agent Systems*, volume 155 of *Advances in Intelligent and Soft Computing*, pages 261–264. Springer Berlin / Heidelberg, 2012.

H. Lopes Cardoso, J. Urbano, A. Rocha, A. J. M. Castro, and E. Oliveira. Ante: Agreement negotiation in normative and trust-enabled environments. In S. Ossowski, editor, *Agreement Technologies*, volume 8 of *Law, Governance and Technology Series*. Springer, 2013.

N. Luhmann. *Trust and Power.* John Wiley & Sons, New York, 1979.

M. W. Macy and Y. Sato. Trust, cooperation, and market formation in the u.s. and japan. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 3):7214–7220, 2002.

M. W. Macy and J. Skvoretz. The evolution of trust and cooperation between strangers: A computational model. *American Sociological Review*, 63(5):638–660, Oct. 1998.

S. Marsh. *Formalising Trust as a Computational Concept.* PhD thesis, University of Stirling, 1994.

S. Marsh and P. Briggs. Examining trust, forgiveness and regret as computational concepts. In J. Golbeck, editor, *Computing with Social Trust*, Human-Computer Interaction Series, pages 9–43. Springer London, 2009.

E. M. Maximilien and M. P. Singh. Agent-based trust model involving multiple qualities. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, July 2005.

R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709–734, July 1995.

M. E. McCullough and W. T. Hoyt. Transgression-related motivational dispositions: Personality substrates of forgiveness and their links to the big five. *Personality and Social Psychology Bulletin*, 28(11):1556–1573, 2002. doi: 10.1177/014616702237583.

D. Melaye and Y. Demazeau. Bayesian dynamic trust model. In M. Pechoucek, P. Petta, and L. Varga, editors, *Multi-Agent Systems and Applications IV*, volume 3690 of *LNCS*, pages 480–489. Springer Berlin / Heidelberg, 2005.

Merriam-Webster.com. January 2012. URL `http://www.merriam-webster.com/`. Web.

J. Mundinger and J.-Y. L. Boudec. Analysis of a reputation system for mobile ad-hoc networks with liars. *Performance Evaluation*, 65(3Ű4): 212–226, 2008. ISSN 0166-5316.

M. Nakatsuji, Y. Fujiwara, A. Tanaka, T. Uchiyama, and T. Ishida. Recommendations over domain specific user graphs. In *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 607–612, Amsterdam, The Netherlands, The Netherlands, 2010. IOS Press. ISBN 978-1-60750-605-8.

R. Neisse, M. Wegdam, M. Van Sinderen, and G. Lenzini. Trust management model and architecture for context-aware service platforms. In *Proceedings of the 2007 OTM confederated international conference on On the move to meaningful internet systems: CoopIS, DOA, ODBASE, GADA, and IS - Volume Part II*, OTM'07, pages 1803–1820, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3-540-76835-1, 978-3-540-76835-7.

C. T. Nguyên and O. Camp. Using context information to improve computation of trust in ad hoc networks. In *Proceedings of the 2008 IEEE*

International Conference on Wireless & Mobile Computing, Networking & Communication, WIMOB '08, pages 619–624, Washington, DC, USA, 2008. IEEE Computer Society. ISBN 978-0-7695-3393-3. doi: 10.1109/WiMob.2008.43.

E. Oliveira. A structured environment to facilitate agreements. In G. V. S. Ossowski, F. Toni, editor, *Proceedings of the First International Conference on Agreement Technologies (AT 2012)*, volume 918, pages 351–352, Dubrovnik, Croatia, 2012. CEUR Workshop Proceedings.

E. Oliveira and A. P. Rocha. Agents advanced features for negotiation in electronic commerce and virtual organisations formation processes. In *Agent Mediated Electronic Commerce, The European AgentLink Perspective.*, pages 78–97, London, UK, 2001. Springer-Verlag. ISBN 3-540-41671-4.

B. Padovan, S. Sackmann, T. Eymann, and I. Pippow. A prototype for an agent-based secure electronic marketplace including reputation-tracking mechanisms. *Int. J. Electron. Commerce*, 6:93–113, July 2002. ISSN 1086-4415.

F. Paglieri and C. Castelfranchi. Trust in relevance. In S. Ossowski, F. Toni, and G. A. Vouros, editors, *Proceedings of the First International Conference on Agreement Technologies, AT 2012, Dubrovnik, Croatia*, volume 918, pages 332–346. CEUR-WS.org, 2012.

G. Paliouras, V. Karkaletisis, C. Papatheodorou, and C. D. Spyropoulos. Exploiting learning techniques for the acquisition of user stereotypes and communities. In *Proceedings of the seventh international conference on User modeling*, UM '99, pages 169–178, Secaucus, NJ, USA, 1999. Springer-Verlag New York, Inc. ISBN 3-211-83151-7.

M. Paolucci and R. Conte. Reputation: Social transmission for partner selection. In G. Trajkovski and S. G. Collins, editors, *Handbook of Research on Agent-Based Societies: Social and Cultural Interactions*, pages 243–260. IGI Global, 2009.

P. Pasquier, R. A. Flores, and B. Chaib-Draa. Modelling flexible social commitments and their enforcement. In M.-P. Gleizes, A. Omicini, and F. Zambonelli, editors, *Engineering Societies in the Agents World V*, volume 3451 of *Lecture Notes in Artificial Intelligence*, pages 139–151. Springer, Toulouse, France, 2005.

J. Patel. *A Trust and Reputation Model for Agent-Based Virtual Organisations*. PhD thesis, University of Southampton, 2006.

P. A. Pavlou, Y.-H. Tan, and D. Gefen. The Transitional Role of Institutional Trust in Online Interorganizational Relationships. *Hawaii International Conference on System Sciences*, 7:215–224, 2003.

I. Pinyol and J. Sabater-Mir. Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review*, pages 1–25, 2011. ISSN 0269-2821.

S. M. Platek, A. L. Krill, and B. Wilson. Implicit trustworthiness ratings of self-resembling faces activate brain centers involved in reward. *Neuropsychologia*, 47(1):289 – 293, 2009. ISSN 0028-3932.

W. Poortinga and N. F. Pidgeon. Trust, the asymmetry principle, and the role of prior beliefs. *Risk Analysis*, 24(6):1475–1486, 2004. ISSN 1539-6924.

J. R. Quinlan. Induction of Decision Trees. *Mach. Learn.*, 1:81–106, March 1986. ISSN 0885-6125.

S. D. Ramchurn, N. R. Jennings, C. Sierra, and L. Godo. Devising a trust model for multi-agent interactions using confidence and reputation. *Applied Artificial Intelligence*, 18(9-10):833–852, 2004.

S. Reece, S. Roberts, A. Rogers, and N. R. Jennings. A multi-dimensional trust model for heterogeneous contract observations. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 1*, AAAI'07, pages 128–135. AAAI Press, 2007a. ISBN 978-1-57735-323-2.

S. Reece, A. Rogers, S. Roberts, and N. R. Jennings. Rumours and reputation: evaluating multi-dimensional trust within a decentralised reputation system. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, AAMAS '07, pages 1–8, New York, NY, USA, 2007b. ACM. ISBN 978-81-904262-7-5.

M. Rehák and M. Pěchouček. Trust modeling with context representation and generalized identities. In *Proceedings of the 11th international workshop on Cooperative Information Agents XI*, CIA '07, pages 298–312, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-75118-2. doi: 10.1007/978-3-540-75119-9_21.

M. Rehak, M. Gregor, and M. Pechoucek. Multidimensional context representations for situational trust. In *Proceedings of the IEEE Workshop on Distributed Intelligent Systems: Collective Intelligence and Its Applications*, DIS '06, pages 315–320, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2589-X.

M. Rehák, M. Pěchouček, M. Grill, and K. Bartos. Trust-based classifier combination for network anomaly detection. In *Proceedings of the 12th international workshop on Cooperative Information Agents XII*, CIA '08, pages 116–130, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-85833-1. doi: 10.1007/978-3-540-85834-8_11.

S. Roccas, L. Sagiv, S. H. Schwartz, and A. Knafo. The big five personality factors and personal values. *Personal. Soc. Psychol. Bull.*, 28(6):789–801, 2002. doi: 10.1177/0146167202289008.

A. Rocha and E. Oliveira. Electronic institutions as a framework for agentsŠ negotiation and mutual commitment. In P. Brazdil and A. Jorge, editors, *Progress in Artificial Intelligence*, volume 2258 of *Lecture Notes in Computer Science*, pages 3–25. Springer Berlin / Heidelberg, 2001. ISBN 978-3-540-43030-8.

A. Rocha, H. Lopes Cardoso, and E. Oliveira. *Virtual Enterprise Integration: Technological and Organizational Perspectives*, chapter Contributions to an Electronic Institution supporting Virtual Enterprises' life cycle, chapter XI, pages 229–246. Idea Group Inc., 2005. ISBN 1-59140-406-1.

A. P. Rocha and E. Oliveira. An electronic market architecture for the formation of virtual enterprises. In *Proceedings of the IFIP TC5 WG5.3 / PRODNET Working Conference on Infrastructures for Virtual Enterprises: Networking Industrial Enterprises*, pages 421–432, Deventer, The Netherlands, The Netherlands, 1999. Kluwer, B.V. ISBN 0-7923-8639-6.

J. B. Rotter. A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35(4):651–665, 1967.

S. Ruohomaa and L. Kutvonen. Making multi-dimensional trust decisions on inter-enterprise collaborations. In *Proceedings of the 2008 Third International Conference on Availability, Reliability and Security*, ARES '08, pages 873–880, Washington, DC, USA, 2008. IEEE Computer Society. ISBN 978-0-7695-3102-1.

J. Sabater. *Trust and Reputation for Agent Societies*. Number 20 in Monografies de l'institut d'investigació en intelligència artificial. IIIA-CSIC, 2003.

J. Sabater and C. Sierra. Regret: Reputation in gregarious societies. In *Proceedings of the Fifth International Conference on Autonomous Agents*, AGENTS '01, pages 194–195, New York, NY, USA, 2001. ACM. ISBN 1-58113-326-X.

J. Sabater-Mir and M. Paolucci. On representation and aggregation of social evaluations in computational trust and reputation models. *Int. J. Approx. Reasoning*, 46(3):458–483, 2007.

J. Sabater-Mir, M. Paolucci, and R. Conte. Repage: Reputation and image among limited autonomous partners. *Journal of Artificial Societies and Social Simulation*, 9(2):3, 2006. ISSN 1460-7425.

M. Sako. Does trust improve business performance? In C. Lane and R. Bachmann, editors, *Trust within and between Organizations: Conceptual Issues and Empirical Applications*. Oxford University Press, 1998.

M. Sako. Does trust improve business performance?, 2002.

A. D. Salvatore, I. Pinyol, M. Paolucci, and J. Sabater-mir. Grounding reputation experiments. a replication of a simple market with image exchange. In *In Proceedings of the M2MŠ07*, pages 32–45, 2007.

F. D. Schoorman, R. C. Mayer, and J. H. Davis. An integrative model of organizational trust: Past, present, and future. *Academy of Management Review*, 32(2):344–354, 2007.

C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, 2001.

P. Slovic. Perceived risk, trust, and democracy. *Risk Analysis*, 13(6):675–682, 1993.

S. Srivastava, O. P. John, S. D. Gosling, and J. Potter. Development of personality in early and middle adulthood: Set like plaster or persistent change? *J. Personal. Soc. Psychol.*, 84(5):1041–1053, May 2003. doi: 10.1037/0022-3514.84.5.1041.

D. Straker. *Changing Minds: in Detail*. Syque Press, Crowthorne, 2008.

T. Strang, C. Linnhoff-Popien, and K. Frank. Cool: A context ontology language to enable contextual interoperability. In J.-B. Stefani, I. Demeure, and D. Hagimont, editors, *Distributed Applications and Interoperable Systems*, volume 2893 of *Lecture Notes in Computer Science*, pages 236–247. Springer Berlin / Heidelberg, 2003. ISBN 978-3-540-20529-6.

Y.-H. Tan and W. Thoen. An outline of a trust model for electronic commerce. *Applied Artificial Intelligence*, 14(8):849–862, 2000.

M. Tavakolifard. Situation-aware trust management. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 413–416, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-435-5. doi: 10.1145/1639714.1639802.

M. Tavakolifard, S. J. Knapskog, and P. Herrmann. Trust transferability among similar contexts. In *Proceedings of the 4th ACM symposium on QoS and security for wireless and mobile networks*, Q2SWinet '08, pages 91–97, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-237-5. doi: 10.1145/1454586.1454603.

M. Tavakolifard, P. Herrmann, and P. Öztürk. Analogical trust reasoning. In E. Ferrari, N. Li, E. Bertino, and Y. Karabulut, editors, *Trust Management III*, volume 300 of *IFIP Advances in Information and Communication Technology*, pages 149–163. Springer Boston, 2009. ISBN 978-3-642-02055-1.

W. Teacy, J. Patel, N. Jennings, and M. Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12:183–198, 2006. ISSN 1387-2532.

W. L. Teacy, M. Luck, A. Rogers, and N. R. Jennings. An efficient and versatile approach to trust and reputation using hierarchical bayesian modelling. *Artificial Intelligence*, 193:149 – 185, 2012. ISSN 0004-3702.

S. Toivonen and G. Denker. The impact of context on the trustworthiness of communication: An ontological approach. In *ISWC Workshop on Trust, Security, and Reputation on the Semantic Web*, volume 127. CEUR, 2004.

N. Tokatli. Global sourcing: insights from the global clothing industry – the case of zara, a fast fashion retailer. *Journal of Economic Geography*, 2007.

J. Tullberg. Trust – the importance of trustfulness versus trustworthiness. *Journal of Socio-Economics*, 37(5):2059–2071, 2008. ISSN 1053-5357. doi: 10.1016/j.socec.2007.10.004.

J. Urbano, A. P. Rocha, and E. Oliveira. Computing confidence values: Does trust dynamics matter? In *Proceedings of the 14th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, EPIA '09, pages 520–531, Berlin, Heidelberg, 2009a. Springer-Verlag. ISBN 978-3-642-04685-8.

J. Urbano, A. P. Rocha, and E. Oliveira. Trust evaluation for reliable electronic transactions between business partners. In *Proceedings of The AAMASŠ09 Workshop on Agent-based Technologies and applications for enterprise interOPerability, Budapest, Hungary, May 12*, pages 85–96, 2009b.

J. Urbano, H. Lopes Cardoso, and E. Oliveira. Making electronic contracting operational and trustworthy. In *12th Ibero-American Conference on Artificial Intelligence*, Bahia Blanca, Argentina, 2010a. Springer.

J. Urbano, A. Rocha, and E. Oliveira. Trustworthiness tendency incremental extraction using information gain. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 411–414, Los Alamitos, CA, USA, Sept. 2010b. IEEE Computer Society.

J. Urbano, H. Lopes Cardoso, E. Oliveira, and A. P. Rocha. Normative and trust-based systems as enabler technologies for automated negotiation. In F. Lopes and H. Coelho, editors, *Negotiation and Argumentation in Multi-Agent Systems (MAS)*. Bentham Science Publishers Ltd., 2011a.

J. Urbano, A. Rocha, and E. Oliveira. A dynamic agentsŠ behavior model for computational trust. In L. Antunes and H. Pinto, editors, *Progress in Artificial Intelligence*, volume 7026 of *Lecture Notes in Computer Science*, pages 536–550. Springer Berlin / Heidelberg, 2011b. ISBN 978-3-642-24768-2.

J. Urbano, A. Rocha, and E. Oliveira. A situation-aware computational trust model for selecting partners. In N. Nguyen, editor, *Transactions on Computational Collective Intelligence V*, volume 6910 of *Lecture Notes in Computer Science*, pages 84–105. Springer Berlin Heidelberg, 2011c. ISBN 978-3-642-24015-7.

J. Urbano, A. P. Rocha, and E. Oliveira. Extracting trustworthiness tendencies using the frequency increase metric. In J. Filipe, J. Cordeiro, W. Aalst, J. Mylopoulos, M. Rosemann, M. J. Shaw, and C. Szyperski, editors, *Enterprise Information Systems*, volume 73 of *Lecture Notes in Business Information Processing*, pages 208–221. Springer Berlin Heidelberg, 2011d. ISBN 978-3-642-19802-1.

J. Urbano, A. P. Rocha, and E. Oliveira. Trust evaluation for reliable electronic transactions between business partners. In K. Fischer, J. P. Müller, R. Levy, W. Aalst, J. Mylopoulos, M. Rosemann, M. J. Shaw, and C. Szyperski, editors, *Agent-based Technologies and Applications for Enterprise Interoperability*, volume 98 of *Lecture Notes in Business Information Processing*, pages 219–237. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-28563-9.

M. Venanzi, M. Piunti, R. Falcone, and C. Castelfranchi. Facing openness with socio-cognitive trust and categories. In T. Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 400–405. AAAI Press, 2011.

D. Villatoro, S. Sen, and J. Sabater-Mir. Of social norms and sanctioning: A game theoretical overview. *International Journal of Agent Technologies and Systems*, 2(1):1–15, 2010.

K. H. Wathne and J. B. Heide. Opportunism in interfirm relationships: Forms, outcomes, and solutions. *The Journal of Marketing*, 64(4):36–51, October 2000.

O. E. Williamson. Transaction-cost economics: The governance of contractual relations. *Journal of Law and Economics*, 22:233–261, October 1979.

Y. Xie and S. Peng. How to repair customer trust after negative publicity: The roles of competence, integrity, benevolence, and forgiveness. *Psychology and Marketing*, 26(7):572–589, 2009. ISSN 1520-6793.

T. Yamagishi and M. Yamagishi. Trust and commitment in the united states and japan. *Motivation and Emotion*, 18:129–166, 1994. ISSN 0146-7239.

B. Yu and M. P. Singh. An Evidential Model of Distributed Reputation Management. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: part 1*, AAMAS '02, pages 294–301, 2002. ISBN 1-58113-480-0.

G. Zacharia and P. Maes. Trust management through reputation mechanisms. *Applied Artificial Intelligence*, 14(9):881–907, 2000.