Chapter 1

Introduction

1.1 Motivation

The importance of reputation and trust is out of question in both human and virtual societies. The sociologist Luhmann wrote [Luhmann, 1979]: *"Trust and trustworthiness are necessary in our everyday life. It is part of the glue that holds our society together"*. Luhmann's observation was also contrasted in virtual societies. The proliferation of electronic commerce sites started the need for mechanisms that ensure and enforce *normative* behaviors and at the same time, increase electronic transactions by promoting potential users' *trust* towards the system and the business agencies (agents) that operate in the site.

Along with it, reputation arises as a key component of trust, becoming an implicit social control artifact [Conte and Paolucci, 2002]. Humans rely on reputation information to *choose* partners to cooperate with, to trade, to form coalitions etc. and it has been studied from different perspectives, such as psychology (Bromley [Bromley, 1993], Karlins *et al.* [Karlins and Abelson, 1970]), sociology (Buskens [Buskens, 1998]), philosophy (Plato [Plato, 1955], Hume [Hume, 1975]) and economy (Marimon *et al.* [Marimon et al., 2000], Celentani *et al.* [Celentani et al., 1966]). Every society has its own rules and norms that members should follow to achieve a *wellfare* society. The social control that reputation generates emerges implicitly in the society, since non-normative behaviors will tend to generate bad reputation that agents will take into account when selecting their partners, and therefore it can cause exclusion due to social rejection.

One of fields that most is using these concepts is the field of multi-agent systems (MAS). These systems are traditionally composed of discrete unites called agents that are autonomous and that need to interact to each other to achieve their goals. The parallelism with human societies is obvious, and also the problems, specially when we are talking about *open* MAS. The main feature that characterizes open multi-agent systems is that the intentions of the agents are unknown. Hence, due to the uncertainty of their potential behavior

we need mechanisms to control the interactions among the agents, and protect *good* agents from fraudulent entities. Traditionally, three approaches have been followed to solve such problems:

- **Security Approach:** At this level, basic structural properties are guaranteed, like authenticity and integrity of messages, privacy, agents' identities, etc. They can be secured by means of cryptography, digital signatures, electronic certificates etc. However, this approach does not tell anything about the quality of the information, although the established control is more than valuable.

- **Institutional Approach:** This approach assumes a central authority that observes, controls or enforces agents' actions, and might punish them in case of *non-desirable* behaviors. It is indisputable that this approach ensures a high control in the interactions, but it requires a centralized hub. Moreover, the control is bounded to structural aspects of the interactions: allowed, forbidden, obliged actions can be checked and controlled. However, the quality of the interactions is left apart, in part, because a *good* or *bad* interaction has a subjective connotation that can depend on the current goals of each individual agent.

- **Social Approach:** Reputation and trust mechanisms are placed at this level. In this approach agents themselves are capable of punishing non-desirable behaviors, y for instance, not selecting certain partners. To achieve such distributed control agents must model other agents' behaviors, and following the similitude with human societies, trust and reputation mechanism arise as a good solution. This requires however the development of computational models of trust and reputation, which must cover not only the generation of social evaluations in all the dimensions, but on dealing with how agents use reputation information to select partners, how agents communicate and spread reputation, and how agents handle communicated reputation information, etc. It is important to remark that these approaches are complementary and that each one covers a different typology of problems, all related to the control of interactions in open MAS.

This work is framed in the field of computational reputation and trust models for open MAS. In the recent years, the scientific research in this field has considerably increased, and in fact, reputation and trust mechanisms have been already considered a key elements in the design of MAS [Luck et al., 2005]. Nowadays, most of the computational models use game theoretical approaches that suffice for simple environments. However, if we want to undertake problems found in socially complex virtual societies, more sophisticated trust and reputation systems based on solid cognitive theories are needed.

Taking the cognitive theory of reputation developed by Conte and Paolucci [Conte and Paolucci, 2002] as a base, we deal with problems that traditionally have been left apart when facing such complex systems. On the one hand, we

deal with pragmatic-strategic decisions by defining an agent architecture capable of integrating reputation information into its deliberative process. On the other hand, we face memetic decisions by specifying a family of argumentation-based dialog protocols that allows the agents to analyze the internal elements used to infer reputation-related concepts, and exchange them with other agents. In the next section we detail the scientific contributions.

1.2 Main Contributions

This work contributes to the field of computational trust and reputation for multiagent systems in three lines:

First - An Ontology of Reputation and the L_{rep} Language

We present an ontology of reputation and the language L_{rep} to capture the reputation information that computational trust and reputation models manage in terms of social evaluations (evaluations about the social performance of an entity in a specific context). It serves to precisely determine the elements that compose a social evaluation and at the same time, provides a clear conceptualization of the involving terms. The main features are:

- The ontology considers computational aspects, such like the representation type used to evaluate other agents performances. For instance, some models use a set of linguistics labels like *Very Bad, Bad, Neutral, Good, Very Good*, while others use probabilistic distributions. We propose four types of representations that capture most of the representations used in the current state-of-the-art models, and define transformation functions to move from one type to another.

- The ontology introduces a taxonomy of social evaluations extracted mainly from the cognitive theory of reputation by Conte and Paolucci [Conte and Paolucci, 2002]. Even when the specific terms may not have a direct connection with the terminology used by other reputation models, the information that most of the current models manage fits into the terms of the ontology.

- The ontology serves as a base to define L_{rep}, a many-sorted first-order language that we use to characterize the reputation information that agents hold. We assume that agents use L_{rep} to write and reason about reputation concepts and associate an inference relation $\vdash_i$ that represents a particular reputation model. With it, we can formalize the fact that even when agents use the same language to express reputation concepts, agents can infer them in multiple and different ways.

Second - The BDI+Repage Model

We introduce the BDI+Repage [Pinyol and Sabater-Mir, 2009a] agent architecture, a *belief-desire-intention* (BDI) architecture that integrates the information that the computational reputation system Repage [Sabater-Mir et al., 2006] provides into the practical reasoning process of the agent. Differently from most of the current state-of-the-art systems that focus on epistemic aspects (how evaluations are calculated), our model deals mainly with the pragmatic aspects of reputation information. The main characteristics of the system are:

- It is modular. The model is defined as a multi-context system (MCS) [Giunchiglia and Serafini, 1994], a framework that allows several distinct theoretical components to be specified together, with a mechanism to relate these components. From a software engineering perspective, MCS supports modular architectures and encapsulation. From a logical modeling perspective, it allows the construction of agents with different and well-defined logics, keeping all formulas of the same logic in their corresponding context. This increases considerably the representation power of logical agents, and at the same time, simplifies their conceptualization. In our model, each main attitude (Belief, Desire and Intention) is specified as an independent context. Also, the Repage system is introduced as a context. Our model specifies then how such contexts are related to each other, defining the practical reasoning path of the agent. This modular architecture permits easy integrations of possible modules that could extend the functionalities of the original one.

- It is based on solid logical frameworks. We use an existing complete logic of preferences based on Lukasiewicz [Casali, 2008] to model desires and intentions, and we introduce a new logic to deal with the beliefs of the agent. The belief logic is a classical first-oder many-sorted logic, deals with probabilities and is capable of representing and combine the information that the reputation model Repage computes. Differently from other probabilistic logics, it handles multiple probability distributions under some restrictive settings, and because it is specified as a first-order logic, it permits a smooth implementation.

- It handles *image* and *reputation*. The Repage model is based on a cognitive theory of reputation that states a main difference between image and reputation. While both objects are social evaluations, image refers to a simple evaluative belief that tells how agents are in a certain context, and reputation is a metabelief, telling that a given social evaluation circulates in the society. The belief logic that we develop captures both concepts and combine them, defining a family of agents depending on how such combination is performed.

- It can be seen as an instantiation of a cognitive trust model. Some cognitive theories of trust suggest that trust is a mental state composed of a set

of beliefs and goals that describe the *decision* to rely on someone, so, it is the result of a practical reasoning process. Our model fits into this description and becomes, as far as we know, the only cognitive trust model that describes each step of the reasoning process.

- It is generic. The model is not attached to any specific domain ontology nor network typology, and inherits the properties and characteristics of the underling reputation model. We use Repage as a paradigmatic example, but any model whose information can be captured by the reputation language L_{rep} could be placed into the system.

Third - An Argumentation-Based Protocol for Reputation Exchange

We develop an argumentation-based dialog protocol for the exchange of reputation-related information. Due to the subjectivity of reputation information, a social evaluation totally reliable by an agent A may not be reliable for B, because the bases under which A has inferred the social evaluation cannot be accepted by B. This can happen because agents have different inference rules, have had different experiences, have different goals, etc. When such information is communicated this can become very problematic, specially if the reputation model assigns a reliability measure to the communicated information, because of the reasons above.

The argumentation-based protocol we develop offers a possible solution for this, and can complement already existing methods. We suggest that, in communicated social evaluations, the reliability measure cannot be dependent on the source agent, but must be fully evaluated by the recipient agent accordingly to its own knowledge. Then, taking advantage of the internal structure of reputation-related information, rather than allow only single communications, we allow agents to *justify* their communications following the guidelines of the argumentation-based protocol. Then, the agent can incrementally construct a tree of arguments with their attack relations that can be used to decide on the reliability (and thus acceptance) of a communicated social evaluation. The main characteristics of the system are:

- Only the recipient agent decides about the reliability of a communicated evaluation. This differs from other approaches in which the source agent attaches a reliability measure to the communicated social evaluation. This makes more difficult for dishonest agents to intentionally send fraudulent information, because they must be aware of the knowledge of the recipient and justify the *lie* accordingly.

- It uses argumentation frameworks to give semantics to the dialog. We exploit the L_{rep} language to completely define how arguments are constructed and how arguments influence one another. We instantiate a weighted abstract argument framework to define the acceptability semantics of a communicated social evaluation.

- It handles quantitative and qualitative graded information. One of the main characteristics of reputation information is that it is graded. Nowadays it is strange to find a model that provides crisp evaluations of the agents. For instance, an agent A may be *bad*, *very bad* or *very good* etc. as a car driver, and this has to be taken into account when arguing about evaluations.

- It permits dialogs between parties that use different reputation models. Even when we assume that agents use the same language to talk and reason about reputation information (L_{rep} language), we suppose that they can use different inference rules (different reputation models) without having to exchange the exact rules that each agent uses for the inferences.

Next section provides a detailed explanation of the structure of the book.

1.3 Overview and Structure of the Work

The book is structured in seven chapters and two appendixes:

Chapter 2: We present the theoretical bases of our work and a survey of the most relevant computational trust and reputation models that currently exists in literature. On the one hand, in the first part of the chapter, we introduce the cognitive theory of reputation presented by Conte and Paolucci [Conte and Paolucci, 2002], relating their definition of *image* and *reputation* with other definitions and with the notion of *cognitive trust* pointed out by some authors. Furthermore we explain Repage [Sabater-Mir et al., 2006], a computational reputation model based on [Conte and Paolucci, 2002] and explore some of its advantages by detailing empirical results that we obtained through simulations. On the other hand, the second part of the chapter is devoted to a survey of the current state-of-the-art reputation and trust models. We describe three other surveys and examine the different dimensions of analysis that each one of them proposes. At the end of the chapter, we also propose a complementary classification.

Chapter 3: The objective of this chapter is to establish a taxonomy of reputation-related concepts. First, we define an ontology of reputation to explicitly state the elements that according to us, are important in the field. Our ontology has a clear computational perspective and serves as a taxonomy of the concepts that our work uses. Second, we introduce the L_{rep} language, a first-order language to express reputation-related concepts described in the ontology, and that agents use to write statement and reason about reputation information. We provide examples to show how the language captures a wide range of state-of-the-art models, specially the Repage model, which currently is one of the most expressive models.

Chapter 4: We introduce the BC-logic, a belief logic capable of integrating reputation information coming from reputation models like Repage, with the

normal beliefs that the agent holds about the domain. It is a sorted first-order logic that manages probability predicates and that subsumes all possible inconsistencies in terms of probabilities. We prof that the proposed theory used by the agents to reason is consistent and decidable, since it can be seen as a set of universal horn clauses.

Chapter 5: The chapter proposes the BDI+Repage architecture. We specify the architecture using multicontext systems [Giunchiglia and Serafini, 1994] and use the logic defined in chapter 4 to manage the belief base of the agent. We specify one context for each main attitude of the agent (Belief, Desire Intention) and design the links (bridge rules) among those contexts, designing a practical reasoning process. Even when we use the Repage reputation model in the integration, it should be taken as a paradigmatic example, since the only requirement is to manage reputation models whose information can be captured by the language L_{rep}.

Chapter 6: While the previous chapter deals with pragmatic-strategic decisions, on how agents use reputation information to decide what to do, this chapter struggles with memetic decisions. We face a particular problem attached to the fact that reputation information is subjective. We define a protocol specifically designed for the exchange of reputation-related information between two-parties that uses argumentation techniques. We exploit the L_{rep} language and use it to build an argumentation system capable of providing a semantics to decide whether a communicated social evaluation can be considered reliable for the agents.

Chapter 7: We conclude our analysis and provide some future research lines.

Appendix A: In this appendix we introduce the concept of conversion uncertainty (CU), a measure of information loss produced when transforming from one representation type to another. We define it as a conditional entropy. We provide the detailed CU calculations for all possible transformations.

Appendix B: We present some implementation details of the BDI+Repage model using the Jason platform [Bordini et al., 2007], and we present preliminary results that empirically proof the differences in behavior of some of the families of agents that we presented in chapter 5.

1.4 Related Publications

The following publications are a direct consequence of the development of this work.

- I. Pinyol, J. Sabater-Mir, P. Dellunde and M. Paolucci. Reputation-Based Decisions for Logic-Based Cognitive Agents. In Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS). In Press.
- I. Pinyol and J. Sabater-Mir. An Argumentation-Based Dialog for Social Evaluations Exchange. In proceedings of the 19th European Conference on Artificial Intelligence (ECAI'10), Lisbon, Portugal. In Press.
- I. Pinyol, R. Centeno, R. Hermoso, V. Torres da Silva and J. Sabater-Mir. Norms Evaluation through Reputation Mechanisms for BDI Agents. In proceedings of the 13th International Congress of the Catalan Association for Artificial Intelligence (CCIA'10). In Press.
- I. Pinyol and J. Sabater-Mir. Metareasoning and Social Evaluations in Cognitive Agents. In Autonomic Computing and Communication Systems,volume 23 of LNISCT, pages 220-235. Springer Berlin Heidelberg, 2010.
- I. Pinyol and J. Sabater-Mir. Pragmatic-strategic reputation-based decisions in bdi agents. In proceedings of the 8th International Conference on Autonomous Agents and Multiagents Systems (AAMAS'09), Budapest, Hungary,pages 1001-1008, 2009.
- I. Brito, I. Pinyol, D. Villatoro, and J. Sabater-Mir. HIHEREI: human interaction within hybrid environments regulated through electronic institutions. In proceedings of the 8th International Conference on Autonomous Agents and Multiagents Systems (AAMAS'09), pages 1417-1418, 2009 (best student demo award),
- I. Pinyol and J. Sabater-Mir. Towards the definition of an argumentation framework using reputation information. In Proceedings of the workshop Trust in Agent Societies (TRUST@AAMAS09), Budapest, Hungary., pages 92-103, 2009.
- S. Konig, I. Pinyol, D. Villatoro, J. Sabater-Mir, and T. Eymann. An architecture for simulating Internet-of-services economies. In Proceedings of 7th German conference on Multiagent System Technologies (MATES'09), Hamburg, Germany. Volume 5774 of LNCS Science. Springer, 2009.
- I. Pinyol, J. Sabater-Mir, and P. Dellunde. Probabilistic dynamic belief logic for image and reputation. Frontiers in Artificial Intelligence and Applications, IOS Press, vol 184 p197-205, Spain, 2008.
- I. Pinyol and J. Sabater-Mir. Cognitive social evaluations for multi-context bdi agents. In 9th Annual International Workshop Engineering Societies in the Agents World (ESAW'08), 2008.
- I. Pinyol and J. Sabater-Mir. Arguing about reputation. the lrep language. In 8th Annual International Workshop Engineering Societies in the Agents World, volume 4995 of LNCS, pages 284-299. Springer, 2007.

- I. Pinyol, M. Paolucci, J. Sabater-Mir, and R. Conte. Beyond accuracy.Reputation for partner selection with lies and retaliation. Multiagent-based Simulation (MABS'07). Volume 5003 of LNCS, pages 128-140.Springer, 2007.
- I. Pinyol, J. Sabater-Mir, and G. Cuni. How to talk about reputation using a common ontology: From definition to implementation. In Proceedings of the Ninth Workshop on Trust in Agent Societies (TRUST@AAMAS'07). Hawaii, USA., pages 90-01, 2007.
- A. di Salvatore, I. Pinyol, M. Paolucci, and J. Sabater. Grounding reputation experiments. a replication of a simple market with image exchange. In Proceedings of the Model to Model Workshop (M2M07), Marseille, France, pages 32-45, 2007.
- J. Sabater-Mir, I. Pinyol, D. Villatoro and G. Cun. Towards Hybrid Experiments on Reputation Mechanisms: BDI Agents and Humans in Electronic Institutions. Proceedings of the 12th Conference of the Spanish Association for Artificial Intelligence (CAEPIA'07),Salamanca, SPAIN, 2007.