Building robust trust model for multi-agent systems

Weng, Jian Shu

2008

Weng, J. S. (2008). Building robust trust model for multi-agent systems. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/13588>

<https://doi.org/10.32657/10356/13588>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**BUILDING ROBUST TRUST MODEL FOR
MULTI-AGENT SYSTEMS**

**WENG JIANSHU
SCHOOL OF COMPUTER ENGINEERING
2008**

Building Robust Trust Model for Multi-agent Systems

WENG JIANSHU

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Doctor of Philosophy

2008

Abstract

In multi-agent systems, modeling of trust equips agents with the ability to establish trust in one another based on their personal previous interaction experiences. In the case that agents' personal interaction experiences are not sufficient to establish trust, third-party testimonies are usually sought and aggregated. However, the presence of *unfair testimonies* deteriorates the performance of trust models. Unfair testimonies are testimonies deviating from agents' real behaviors that are perceived by the testimony receivers. The mitigation of the adverse effects of unfair testimonies is a fundamental problem in the research on trust modeling in multi-agent systems.

This thesis presents the research into making trust models robust in the presence of unfair testimonies. It proposes an uncertainty-based testimony-filtering method to mitigate the adverse effects of unfair testimonies. The proposed filtering method is then revamped with a novel credibility model. The revamped model not only improves the effectiveness of the filtering method, but is also general enough to be applied to the existing trust models. Unlike most of the existing methods, the proposed credibility model does not require any additional mechanism or knowledge other than the testimonies shared among the agents. Empirical evaluations also show that our proposal consistently outperforms related work.

This work also goes beyond the existing methods in that it discards the common assumption of the existing methods that each agent would obtain testimonies from all other agents directly in the system. Instead, we propose a credibility-aware referral process on top of the credibility model. The credibility-aware referral process facilitates agents' testimony discovery in an efficient manner, in which more credible agents are iteratively requested to discover testimonies by testimony discovery initiator. Furthermore, this thesis proposes an approach to counteract malicious referrers during testimony discovery. The presence of malicious referrers aggravates the adverse effects of unfair testimonies. This effect is simply ignored in the existing methods.

The proposed research can be applied in many application domains. Some typical domains include (1) e-commerce systems such as the online auction website `ebay.com` where people can transact with other people geographically located thousands of miles away, (2) Peer-to-Peer content sharing networks with which digital contents can easily be shared among users widely distributed in different locations around the world, (3) the Grid with which various computing resources can be shared among various users and institutions, and (4) massive multi-player online role playing games (MMORPGs) in which various user avatars interact with each other.

Acknowledgments

This thesis would not be possible without the help of many people.

First and foremost, I would like to express my gratitude to Assistant Professor **Chunyan Miao**, my supervisor, for her invaluable guidance and persistent support for my research. I also owe a lot to Professor **Angela Goh**, as my co-supervisor, who has contributed so many invaluable suggestions for my research. I would also like to express my gratitude to her for her warm encourage for my research.

I am deeply in debt to the love and support of my family. I would like to give my deepest appreciation to my fiancée, **Yijing Lin**. She is the greatest source of comfort and encouragement for me. Without her support and encouragement, I will not be able to continue my Ph.D study. Appreciations are also given to my grandparents and sisters for their persistent support. I would never go this far without their support.

Thanks also go to School of Computer Engineering, Nanyang Technological University for providing scholarships to support my doctoral study. I would also like to thank administrative and technical staffs of School of Computer Engineering who are always kind enough to help in their respective roles.

Last but not least, I want to thank all my friends, especially Zhuowei Li, Qiankun Zhao, and Baoyao Zhou, for their helps, both in life and research, and also the happiness they have shared with me.

Thank you all!

Contents

Abstract	i
Acknowledgments	iii
List of Figures	ix
List of Tables	xi
1 Introduction	1
1.1 Definition of Trust	1
1.1.1 Trust Information	4
1.2 Trust in Multi-Agent Systems	5
1.2.1 Agents and Multi-Agent Systems	5
1.2.2 Example Scenario of Trust in Multi-Agent Systems	8
1.2.3 Difference between Machine Learning	10
1.3 Research Scope and Motivations	12
1.4 Contributions	16
1.5 Outline of the Thesis	18
2 Literature Review	21
2.1 Factors of a Computational Trust Model	21
2.1.1 Paradigm of the Model	22
2.1.2 Sources of Trust Information	23
2.1.3 Granularity of Trust Information	23
2.1.4 Context-dependence	25
2.1.5 Methods to Tackle Unfair Testimonies	26
2.2 Centralized Computational Trust Models	27

2.2.1	eBay Trust Model	27
2.2.2	Jurca and Faltings	28
2.3	Decentralized Computational Trust Models	29
2.3.1	Models without Method to Tackle the Unfair Testimonies	29
2.3.1.1	FIRE	29
2.3.2	Models with Endogenous Method to Tackle the Unfair Testimonies	31
2.3.2.1	Beta Reputation System	31
2.3.3	Models with Exogenous Method to Tackle the Unfair Testimonies	32
2.3.3.1	Abdul-Rahman and Hailes	32
2.3.3.2	ReGreT	33
2.3.3.3	Yu And Singh	35
2.3.3.4	Kafali and Yolum	36
2.4	Summary	39
3	An Uncertainty-based Method To Filter Unfair Testimonies	43
3.1	Introduction	43
3.2	The Basic Trust Model	44
3.2.1	Notations	44
3.2.2	Maintaining Personal Interaction Experiences	45
3.2.3	Making the Trust Evaluation	46
3.2.4	Aggregating Third-party Testimonies	47
3.3	Extension with the Uncertainty-based Testimonies Filtering	48
3.4	Experimental Study	51
3.4.1	Experimental Methodology	51
3.4.2	The Results	54
3.4.2.1	Comparison with Whitby et al.'s Proposal	56
3.5	Discussion	59

4	Coping with Unfair Testimonies: A General Credibility Model	63
4.1	Introduction	63
4.2	Notations	65
4.3	Maintaining the Witness' Profile	65
4.3.1	Preprocessing the Testimonies	66
4.3.2	The Organization of the Profile	67
4.4	Measuring the Credibility	68
4.4.1	The Derivation of the Credibility Metric	69
4.4.2	Bootstrapping of the Profile and Credibility	70
4.4.3	Evolution of the Credibility	72
4.5	Testimony Filtering and Aggregation	72
4.6	Experimental Study	76
4.6.1	Methodology	77
4.6.2	Results	80
4.7	Summary	84
5	Empirical Evaluation of the Credibility Model	87
5.1	Introduction	87
5.2	Design of TrustBed	88
5.2.1	Interactions among Agents	89
5.2.2	Provider Population Configuration	90
5.2.3	Consumer's Cost and Gain	91
5.2.4	Presence of Unfair Testimonies	92
5.3	Methodology	94
5.3.1	Comparison with Related Work	94
5.3.2	Measurement Metrics	95
5.4	Experimental Setup	96
5.5	Effectiveness in the Presence of Individual Unfair Testimonies	98
5.6	Effectiveness in the Presence of Collusive Unfair Testimonies	103
5.7	Influence of Value of Z	105

5.8	Empirical Evaluation with the ART Testbed	106
5.8.1	Overview	107
5.8.2	Architecture of ART Testbed	108
5.8.3	Results of ART Competition 2006	110
5.9	Summary	112
6	Counteracting the Unfair Referrers: Measuring the Referral-Credibility	115
6.1	Introduction	115
6.2	Credibility-aware Referral Process	117
6.3	Revamping the Referral Process with Modeling of Referral-Credibility	122
6.4	Modeling Agent's Referral-Credibility	123
6.5	Empirical Evaluation	129
6.5.1	Experimental Setup and Methodology	129
6.5.2	The Results	131
6.6	Summary	133
7	Conclusions	137
7.1	Research Contributions	137
7.2	Potential Applications	139
7.2.1	Application in P2P Network and Grid	139
7.2.2	Application in MMORPG	140
7.3	Future Directions	141
	Appendices	144
A	Credibility-modeling Method proposed by Yu and Singh	145
B	List of Publications	147
	Bibliography	149
	Index	155

List of Figures

2.1	Taxonomy of the review trust models	27
3.1	Examples of beta distribution	46
3.2	Uncertainty-based testimonies filtering and aggregation	51
3.3	Adverse effect of the unfair testimonies: (a) without unfair testimonies; (b) with badmouthing testimonies	55
3.4	The trust model in the presence of unfair testimonies	55
3.5	Effectiveness in the presence of ballot-stuffing testimonies	57
3.6	Effectiveness in the presence of badmouthing testimonies	58
4.1	Agent's profile of giving testimonies	67
4.2	Agent a_w 's profile after 3 successful testimony-reporting	69
4.3	A special case leading to 0/0 when evaluating credibility	71
4.4	Agent a_w 's profile after its first successful testimony-reporting	72
4.5	Procedure to update credibility	73
4.6	Agent a_w 's profile after 5 successful testimony-reporting	75
4.7	The credibility model	77
4.8	Comparison of the credibility model and the filtering method in the presence of ballot-stuffing testimonies	81
4.9	Comparison of the credibility model and the filtering method in the presence of badmouthing testimonies	82
5.1	Average gain of each group without unfair testimonies	99
5.2	Average gain of each group in the presence of individual unfair testimonies	100

5.3	Performance of different groups in the presence of various configurations of individual unfair testimonies	101
5.4	Performance of different groups in the presence of various configurations of collusive unfair testimonies: (a) gain; (b) collusion power	104
5.5	Influence of parameter Z	106
6.1	Referral Process	118
6.2	Credibility-aware referral process	120
6.3	Example of referral process	121
6.4	Referral process with the incorporation of referrer's <i>referral-credibility</i>	124
6.5	Examples of the reversed referral chains	126
6.6	Spreading activation based referral-credibility update	128
6.7	Example of referral-credibility update	129
6.8	Performance of different groups in the presence of various combination of malicious referrers and collusive unfair testimonies: (a) 20% malicious referrers; (b) 40% malicious referrers; (c) 60% malicious referrers;	132
6.9	The Complete Trust Model	134

List of Tables

2.1	Computational Trust Models	40
4.1	Comparison of performance without unfair testimonies	83
4.2	POU w.r.t. different ratios of witness giving unfair testimonies	84
5.1	Interaction among Agents in TrustBed	90
5.2	Configurations of providers' willingness	90
5.3	Population of the providers	91
5.4	Experiment parameters	98
5.5	Comparison of performance without unfair testimonies	99
5.6	Performance of each group in the presence of individual unfair testimonies Pos100 100	
5.7	Parameters of ART testbed and competition	109
5.8	Parameter settings of ART competition 2006	110
5.9	Results of ART competition 2006	112

Chapter 1

Introduction

Trust has long been a research topic in various fields of study such as sociology, economics, and philosophy. It has also become an emerging research area in the discipline of computer science [BK02, RHJ04, JIB07]. This chapter gives an introduction to the concept of trust, the extension of trust into the multi-agent systems, the scope of the research documented in this thesis, and the contribution of this thesis.

1.1 Definition of Trust

Trust is an ubiquitous concept that underlies each interaction in our every-day life. For example, a person trusts that his parents will support him, and his friends will be kind to him and will not hurt him. When he transacts with other persons, he trusts that his transaction partners will give him the goods after receiving the payments. He trusts not only individual but also institutions. For example, he can leave his letters and parcels at the post office because he trusts that the post office will send them to the destination in good condition. It is not an exaggeration to say that *trust* is the “glue” that makes interactions between people possible and keeps human society functioning.

As shown above, trust plays an important role in regulating interactions. Nevertheless, trust remains an abstract concept. It is still not clear what trust is and how trust is established or measured. Historically, trust has long been a research topic in various disciplines such as sociology, economics, and philosophy. Since trust is the subject of research in multiple disciplines, it is not surprising to see different working definitions of trust in the literature (see [MC96] for a general

review). A comprehensive discussion of the definitions of trust is beyond the scope of this thesis. Instead, we review a number of well-recognized definitions.

One of the most well recognized definitions is proposed by Deutsch [Deu62] which states that:

“(a) An individual is confronted with an ambiguous path, a path that can lead to an event perceived to be beneficial or to an event perceived to be harmful; (b) he perceives that the occurrence of these events is contingent on the behavior of another person; and (c) he perceives the strength of a harmful event to be greater than the strength of a beneficial event.

If he chooses to take an ambiguous path with such properties, he makes a trusting choice. Otherwise he makes a distrustful choice.”

To have a better understanding of this definition, we revisit the example at the beginning of this chapter.

Example 1.1 *A person is now making the decision whether to transact with another person. The outcome of the transaction can be beneficial or harmful. If his transaction partner defaults the obligations after receiving his payment, the outcome of the transaction is harmful since he receives nothing in return for his payment. On the other hand, the outcome of the transaction would be beneficial since he receives something to compensate his payment. He might gain more if there is value added to the goods. Since his transaction partner has his own interests, and can do anything to maximize his own gain, the transaction partner may be tempted to default in the transaction. Under such a circumstance, if he still opts to complete the transaction, he makes a trusting choice. Otherwise, he makes a distrustful choice.*

This definition describes the nature of *trust*. From the definition itself as well as the above example, it can be seen that trust is primarily about one individual’s decision-making in the presence of *ambiguity* or *uncertainty* of his interaction partner’s behavior. Although Deutsch’s definition answers the question of “what is trust”, it does not explain how trust can be established

(or simply measured). Another definition of trust proposed by Diego Gambetta [Gam88] can be cited to answer this question.

Diego Gambetta's definition [Gam88] interprets trust as

“a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action and in a context in which it affects his own actions”.

According to this definition, the uncertainty, which is in the nature of trust, can be measured as the *expectation* that the interaction partner would take a particular action. For example, the person in the above example can make the trusting choice based on the expectation that his transaction partner would ship the goods. And such an expectation can be measured as the subjective probability that the partner would ship the goods. This definition usually serves as the theoretical foundation of the research on trust models in multi-agent systems (MAS) [RHJ04], whose main focus is to equip the software agents with the ability to quantify the trust they should place in their interaction partners.

People usually interact in multiple contexts. Consequently, the action (as in Gambetta's definition [Gam88]) can be taken in any context. Bernard Barber's definition of trust [Bar83] describes the context of the actions by defining trust as:

- “(a) the expectation of technically competent role performance (e.g. expert knowledge, technical facility, or everyday routine performance);*
- (b) expectations that the natural order (both physical and biological) and the moral social order will persist and be more or less realized;*
- (c) expectations of fiduciary obligation and responsibility, that is, that some others in our social relationships have moral obligations and responsibility to demonstrate a special concern for other's interests above their own.”*

This definition proposes that trust is context-dependent. It can happen in various contexts from the everyday routine performance to fiduciary obligation and responsibility, from natural

order to social order. Moreover, trust in one context does not necessarily imply trust in another context. For example, an individual, say *A*, may trust another individual *B* completely in transactions related to selling breads but may not trust *B* when it comes to transactions of selling cars.

In summary, some common understanding of the concept of *trust* can be observed from the above-mentioned definitions:

- Trust is essentially about one individual's uncertainty relating to another individual's behavior in future interactions.
- The uncertainty about an individual's behavior can be measured as the expectation that it would take a particular action (or behave cooperatively in general).
- Since the action can occur in any context, trust is generally context-dependent. That is, an agent may not necessarily trust another agent with the same degree in different contexts.

1.1.1 Trust Information

Trust information refers to that which can be used by an individual to evaluate the trust it should place in its interaction partner. As discussed, trust is essentially about dealing with the uncertainty relating to the interaction partner's behavior. In this aspect, trust information is sought to reduce the uncertainty about the interaction partner's behavior. Normally, the most direct information available for the trust evaluation is an individual's previous personal interaction experience with a partner. In human society, factors such as the interaction partners' body language or eye contact in physical interactions can also be taken into account.

It is possible that an individual does not have previous interaction experience with the interaction partner or the past interaction experience is not sufficient to reduce the uncertainty about the partner's behavior. In this case, third-party testimonies about their past interaction experience with the same interaction partner would improve the uncertainty in the interaction partner. In other words, the effect of the third-party testimonies is to reduce an individual's uncertainty about its interaction partner's behavior and to make it more certain of its trust in its interaction

partner [BK95]. In human society, people can also rely on reports from formal institutions. For example, in monetary-related transactions, people can rely on reports from third-party credit report bureaus to make trust evaluations.

Since the testimonies are provided on a subjective basis, the testimony-receivers cannot control the sincerity and fairness of the testimonies. There exists the presence of *unfair testimonies*, i.e. testimonies deviating from an individual's real behaviors that are perceived by the testimony-receiver. Given this, the third-party testimonies should not be utilized blindly. Instead, the testimony-receiver should choose different strategies in dealing with the third-party testimonies based on the fairness of the testimonies. The fairness of the testimonies is according to the view of the testimony-receiver. It should be considered as the subjective measurement of the testimony's usefulness, which is interpreted as whether it has reduced the testimony-receiver's uncertainty.

There are also cases that testimonies increase the testimony-receiver's uncertainty about the interaction partners' future behavior. In such cases, the testimonies can also be considered as "useful" in some sense, because they help the testimony-receiver to avoid the corresponding interaction partners due to its increased uncertainty about the interaction partners. Nevertheless, such testimonies do not really contribute to the purpose of trust evaluation, which is to help individuals identify the right interaction partners. Hence, this thesis only considers the testimony that has reduced the testimony-receiver's uncertainty as useful, and measures its usefulness accordingly.

1.2 Trust in Multi-Agent Systems

Trust has also become a relevant research topic in the domain of computer science in recent years, especially multi-agent systems [BK02, RHJ04, JIB07].

1.2.1 Agents and Multi-Agent Systems

First of all, we introduce the concept of *agent* since it is the main building block of a multi-agent system (MAS). A widely recognized definition of agent is [Woo97]:

An agent is an encapsulated computer system situated in some environment and capable of flexible, autonomous action in that environment in order to meet its design objectives.

According to this definition, agents are generally considered to have the following characteristics [WJ95, Jen01]:

- *clearly identifiable problem-solving entities with well-defined boundaries and interfaces;*
- *situated (embedded) in a particular environment over which they have partial control and observability - they receive inputs related to the state of their environment through sensors and they act on the environment through effectors;*
- *designed to fulfill a specific role - they have particular objectives to achieve;*
- *autonomous - they have control both over their internal state and over their own behavior;*
- *capable of exhibiting flexible problem-solving behavior in pursuit of their design objectives - being both reactive (able to respond in a timely fashion to changes that occur in their environment) and proactive (able to opportunistically adopt goals and take the initiative).*

Other than these characteristics, agents are usually assumed to be rational. That is, agents will take whatever action that is expected to maximize their gains. On the other hand, irrational agents have unpredictable behaviors, and are usually excluded from the research in MAS.

Recent years have seen an increasing number of software agents being developed to extend the region of human interaction within environments such as the Internet [HJL03]. Such agents, with their autonomous reasoning and decision-making capability, can engage in complex interactions on behalf of their owners, which can be human or institutions. Agents usually live in a society of agents. Such a society of agents is termed as a *multi-agent system* (MAS) [Woo02]. According to [Syc98], the characteristics of a MAS are that:

(1) each agent has incomplete information or capabilities for solving the problem and, thus, has a limited viewpoint; (2) there is no system global control; (3) data are decentralized; and (4) computation is asynchronous.

In short, a MAS is an open and decentralized computer system. Here, “open” primarily means that (1) participating agents in a MAS are owned by different stakeholders, each of whom has different aims and objectives [RHJ04], and (2) there is no central entity to control participating agents’ behaviors. Within a MAS, several agents interact and work together with one another to solve problems that are beyond the capability of each individual agent [Jen01, JSW98].

Multi-agent system, with its decentralized nature and independent participants’ autonomous reasoning and decision-making capabilities, has emerged as a natural model for large-scale computer systems in recent years [RHJ04]. Many systems that can be modeled as a MAS. Some typical examples include (1) e-commerce systems such as the online auction website `ebay.com` where people can transact with other people who are geographically located thousands of miles away [RZ02], (2) Peer-to-Peer content-sharing networks with which digital contents can be shared among peers widely distributed in different locations around the world [DdVP⁺02], (3) the Grid with which various computing resources can be shared among users [FKT01], and (4) massive multi-player online role playing games (MMORPGs) in which various user avatars interact with each other [Smi04].

Due to the open nature of MAS, there may be unreliable agents that do not fulfill the obligations hitherto agreed with other agents. There may also be agents that intentionally default their interaction partners to maximize their own interests. Hence, there is a risk when an agent engages in interactions with other agents. Moreover, the risk may even be intensified by the characteristics of interactions in a MAS which are unseen in human society. For example, software agents are able to transact without human intervention in a MAS (e.g. the online auction website such as `ebay.com`), which intensifies the risk that agents would default by not shipping the items after receiving the payments [RZ02, Nat05]. Hence, trust, which is essential in regulating the interaction in human society, needs to be maintained to regulate the interactions in online agent-mediated environments [BK02, RHJ04, JIB07]. The next subsection shows a scenario of trust in MAS and discusses different types of trust in MAS.

1.2.2 Example Scenario of Trust in Multi-Agent Systems

In order to show the importance of trust in multi-agent systems, an example scenario¹ is given in this subsection. The scenario, based on an agent-mediated supply chain, outlines the interactions that may be involved in the domain of agent-mediated business process management. The scenario involves different stakeholders (e.g. distributors, sellers and buyers) that are mediated by agents. Agents wrap Web Services, which facilitates the automated business process management. A purchasing agent at the Bob Company (BC in short) broadcasts a solicitation for 100 PCs needed within 2 days. David Company (DC in short), a distributor, has an agent that subscribes to such solicitations. DC's agent then starts to look for possible suppliers:

1. DC's agent is instructed to look for some reliable suppliers. It also has the information about the acceptable cost for this request.
2. DC's agent contacts a directory service to get a list of all the available suppliers.
3. DC's agent forwards the solicitation of BC to all the available suppliers. Agents of those interested suppliers respond with their current inventory levels.
4. For those agents who respond, DC's agent contacts agents of other distributors, with whom it has cooperated before, and asks them to give opinions of the past transaction experiences with those suppliers. DC's agent then collects the opinions, aggregates them with its own past transaction experience to evaluate those suppliers' reliability.
5. DC's agent contacts the one whose current inventory level is able to fulfill the request, and asks for its offer of price. There are also cases that the request must be fulfilled by more than one supplier because any single supplier is not able to fulfill the request. In this case, DC's agent would aggregate the services provided by more than one supplier to fulfill the request. In both cases, DC's agent, as a rational agent, would select those reliable suppliers according to the evaluations in Step 4. At the same time, before engaging in

¹This example scenario is adapted from the vision of agent-powered Web which is set out by Dr. Jay M. Tenenbaum in [Ten05].

further interaction with the supplier(s), DC's agent asks the supplier(s) to provide digital certificate(s) signed by a recognized certificate authority (CA) to show they are really who they claim to be.

6. DC's agent and the agent(s) of the supplier(s) negotiate the details regarding how to fulfill the request, such as the amount and price.
7. DC's agent reaches an agreement with the suppliers, and informs BC's purchasing agent that it can fulfill the request.
8. Upon being informed by DC's agent, BC's agent asks for approval from the manager, whose agent automatically approves the order because it falls within DC's contractual guidelines and credit limit. After that, BC's agent places order with DC's agent.
9. After receiving the order from BC's agent, DC's agent informs BC's agent that it accepts the order. Then BC's agent notifies the financial department to pay DC electronically.

This scenario fits well into the scope of MAS in that the participating agents are owned by different stakeholders (e.g. the agent owned by BC, DC, and those suppliers), and they all have their own objectives. For DC's agent, the environment is full of uncertainty in that it is not sure whether the other agents are reliable. For example, the agents owned by the supplier(s) may not fulfill their obligations. Nevertheless, a transaction is still possible, thanks to the trust among the various agents and/or non-agent entities and institutions. In fact, trust occurs in almost every step of an agent's interaction. For example:

- In Step 1, DC knows the capability of its agent, and trusts that the agent is competent in completing the task of finding suitable supplier(s).
- In Step 2, DC's agent trusts that the directory service would return a list with correct and complete information about the suppliers.
- In Step 4, DC's agent trusts that most of the distributors' agents it contacts would share their opinions honestly.

- In Step 5, DC's agent trusts that the certificate system is robust enough and the supplier's agent will not forge a faked certificate.
- In Step 9, upon accepting the order, DC's agent trusts that the suppliers would not default their obligation specified in the agreement (in Step 7) to fulfill the request. It also trusts that the electronic payment service would really debit credit into the correct accounts.

1.2.3 Difference between Machine Learning

Machine learning is concerned with creating algorithms that can improve their performance at certain task as they acquire experience or data [Mit97]. In general, machine learning algorithms extract rules and patterns out of massive data sets (i.e. training data), and the derived rules and patterns are generalized enough to be applied on future data (i.e. test data) which usually has the similar statistical characteristics as the training data. Common machine learning algorithms can be organized into following types, based on the outcome of the algorithms:

- Supervised learning, in which the training data consists of pairs input and desired output. The algorithms need to generate a function that maps inputs to desired outputs, so that it is able to predict the corresponding output for any new valid input after seeing a number of training examples. The output can be a continuous value (i.e. regression) or a class label (i.e. classification).

There are cases that input data without corresponding output is abundant but those with output are expensive to obtain. In this case, the learning algorithms can actively query for the output. This form of supervised learning is called active learning [CGJ96]. The number of training data needed in active learning is often much lower than that required in normal supervised learning.

- Unsupervised learning, in which the desired outputs of the training data are not available or not used. One typical form of unsupervised learning is clustering, in which similar inputs are grouped together.

- Reinforcement learning algorithms attempt to find a policy that guides the agent's action in different environment states in order to maximize long-term reward [SB98]. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm [SB98].

Nevertheless, there are fundamental differences between trust modeling and machine learning in that:

- As discussed above, machine learning extracts generalized rules and pattern out of a set of training data. Based on different set of training data, different set of rules would be extracted. On the other hand, the primary concern of trust modeling is to design a metric that can be applied by agents to measure their interaction partners' trustworthiness based on past their behaviors. The metric itself does not depend on the agents' behaviors.
- Machine learning concerns about the generalization capability of the extracted rules and patterns. That is, the rules and patterns must be applicable on new data, which is not observed in the training data set. However, in trust modeling, one agent's trustworthiness cannot be used on another agent. Instead, different agents' trustworthiness need to be measure individually based on their behavior, though with same trust metric. Additionally, even for same agent, its trustworthiness will be updated every time a new interaction is finished, instead of using one value (derived based on past interactions) for all new interactions.
- Trust modeling might seem to be related to reinforcement learning: the agents' trustworthiness gets updated every time a new interaction is finished, like what the reinforcement learning algorithm does to revise the policy when a new feedback is received. The attempt of reinforcement learning is to find a policy to guide agent's action in different environment states. However, as a rational agent, each agent's action policy has already been set, i.e. to choose the partner with highest trustworthiness to interact with. Past interactions with other agents help to find the right interaction partner, but not to obtain the action policy.

1.3 Research Scope and Motivations

It is seen from the example scenario in Section 1.2.2 that trust occurs in many steps of an agent's interaction. The occurrence of trust in the example scenario (and MAS in general) can be classified into the following four main categories:

- Trust between agents and their owners. This category of trust is related to the owner's expectation of the agents' capabilities in fulfilling the delegated tasks. For example, when David Company instructs its agent to find suitable suppliers, it trusts that the agent is able to handle this task, and will try the best to maximize its interests. According to Grandison and Sloman's classification of trust [GS00], this category of trust falls in the scope of the *delegation trust*.
- Trust among agents. This category of trust covers an agent's expectation that another agent would take a particular action (or in general, behave cooperatively in MAS). It is established based on agents' past behavior. For example, in Step 9, DC's agent trusts that the suppliers would fulfill its obligations based on their behavior in the past transactions. This category of trust is termed as the *service provision trust* in Grandison and Sloman's classification [GS00], and it basically captures agents' *reliability* in providing certain services.
- Trust on the infrastructure. This category of trust covers the trust that an agent must place in the entities and systems of the infrastructure, which do not act like agents. Examples of such entities and systems include the certification system, the electronic payment system, and the directory service in the example scenario. For instance, DC's agent trusts that the certification system is robust enough to counter the agents' attempts to forge certificates; and it also trusts that the directory service would return a list of service providers with correct and complete information. This category of trust is the *infrastructure trust* in Grandison and Sloman's classification [GS00].
- Trust produced by the ownership of certificates. This category of trust covers the expectation that another agent is able to show the ownership of a digital certificate. It depends on

the trust on the certification system, which is part of the *infrastructure trust*. Generally, this category of trust is related to certificate-based authentication and authorization. It makes an agent believe that it is interacting with some one that the latter claims to be, but not some one else. This category of trust is the *certification trust* in Grandison and Sloman's classification [GS00].

As discussed before, trust is introduced to reduce the risks that agents take when interacting with other agents, and to regulate the interactions among agents in a MAS. Hence, among those categories, this thesis mainly studies the trust between the (software) agents, i.e. the service provision trust. That is, our research focuses on agents' ability to evaluate their interaction partners' trustworthiness, which capture their reliability in providing certain service.

The other categories of trust are considered outside the scope of this thesis because:

- Since agents usually act on behalf of their owners to complete some delegated tasks, the owner's understanding and expectation of the agents' capabilities is a prerequisite for the decision to delegate a task to the agent. Hence, trust between agents and their owners can naturally be assumed. Otherwise, their owners would not delegate the tasks to the agents.
- The infrastructure trust and certificate trust are usually policies implemented by agents' owners. Hence, agents are forced to trust the directory service, the payment system, and the certificate system since they are only acting on behalf of their owners. They cannot change the policies. Furthermore, those policies do not influence the interactions among agents too.

Recent years have seen a significant body of work on trust modeling in MAS. Many computational trust models have been proposed to equip the agents with the ability to evaluate their interaction partners' trustworthiness, i.e. the ability to quantify the service provision trust they should place in their interaction partners (see [RHJ04, JIB07] for a general review).

The measurement of trustworthiness depends on availability of the trust information in MAS. However, the nature of MAS has invalidated some trust information that is usually used in human

society. For example, the interaction partners' body language and eye contact during the course of interaction are no longer available due to the lack of physical contact in the interactions in MAS. Reports from formal institutions are generally unavailable since such kind of central authority is difficult to build and maintain because of the open and decentralized nature of MAS. Nevertheless, the following mechanisms are available in MAS:

- Historical interactions storage for the participating agents.

Agents are designed with the capability to store the historical interactions with other agents, which serves as a repository of previous experiences. Each entry in an agent's memory is basically that agent's feedback of its partner's behavior in one specific interaction.

- Sharing of historical interactions.

Instead of storing historical interactions only for an agent's own use, the agent is able to exploit the communication infrastructure of MAS to share its historical interaction experience with other agents.

The deployment of these two mechanisms has made two sources of trust information available in MAS: (a) the truster agent's private information that is obtained from its past personal interaction experiences with the trustee agent, and (b) testimonies reported by third-parties regarding their past interactions with the same trustee. In this thesis, *truster* is used to refer to the agent who is evaluating the trust it should place in another agent, while that agent whose trustworthiness is being evaluated is called as the *trustee*. The general idea of the computational trust models is that the trustee's past behaviors can be used as evidences to evaluate its trustworthiness. Such evidence is from two sources, i.e. it is either perceived by the truster itself or testified by other agents.

Source (a) is the most reliable information since it is based on an agent's own experience. However, agents' private trust information is insufficient to reduce its uncertainty about the interaction partners' behavior. For example, when an agent just enters a MAS, it would have few interactions with other existing participating agents, or even no interactions. Moreover, during

an agent's participation in a MAS, one would not expect that this agent would meet frequently with other agents [RZ02, BK02]. In all those scenarios, source (a) becomes insufficient due to the lack of interactions. In this case, agents usually aggregate source (b) to collect more evidences. [BK95].

However, a problem arises when third-party testimonies are taken into account. That is, there exists the presence of *unfair testimonies*, i.e. testimonies deviating from the trustee's real behaviors that are perceived by the truster. It is reiterated that unfairness of the testimony is according to the view of the truster. It should be viewed as the subjective measure of the usefulness of the testimonies. Unfair testimonies can happen because of the differing views of the witness (i.e. the agent that reports testimony on the trustee) and the truster. As in the real world, a restaurant may be excellent for a customer, but may be terrible for another. It can also happen because of the differing and conflicting interests due to the fact that agents in a MAS usually represent various stakeholders. Some agents may lie deliberately when they are asked by others for opinions on the trustees, if they can benefit from doing so. In both cases, the unfair testimonies, if combined blindly, can bias the truster's evaluation of trust on the trustee, and deteriorate the performance of the trust models [JIB07, Del00, WJI04].

An example is presented here to show the adverse effect of unfair testimonies. Consider an online transaction website such as eBay.com as an example. A buyer **B** is deciding whether to engage in a transaction with seller **S**. **B** has no previous interaction experiences with **S**. Hence, **B** collects testimonies reported by others², among whom there is a number of **S**'s friends. **S**'s friends intentionally give unfairly positive testimonies. As eBay.com evaluates **S**'s trustworthiness as a summary of all the ratings that **S** has received from the other agents over the past few months (one month, 6 months, and 12 months), these unfairly positive testimonies given by its friends will definitely increase the number of positive ratings that **S** has accumulated. This will consequently assist **S** to gain a trustworthiness higher than it deserves. The manipulated outcome of trust evaluation might lead to **B**'s wrong decision to transact with **S**. This could result in monetary loss of **B** if **S** is a malicious seller who usually defaults its obligations to ship the item after receiving

²In fact, in eBay.com, every agent's testimonies are already shared with the public.

the payments. As this simple example shows, the presence of unfair testimonies does deteriorate the performance of the computational trust models.

One may argue that the trust model used by `eBay.com` is naive so that it is susceptible to the presence of unfair testimonies. However, according to recent research, the presence of the unfair testimonies is a common problem faced by every computational trust model [JIB07]. At the early stage of the research on trust modeling in MAS, researchers primarily focused on setting up infrastructure to make trust information available in MAS and designing models to equip agents with the ability to evaluate trust quantitatively based on the available information. They usually assume that every agent would tell the truth when sharing testimonies, and ignore the presence of unfair testimonies. As MAS continue to grow and mature, they depend more on trust to regulate the interactions. Making the trust models in MAS robust in the presence of unfair testimonies deserves an extensive study. How to mitigate or even avoid the adverse effect of unfair testimonies is a problem that every computational trust model should address. It has also been identified as one of the fundamental problems in the research on trust modeling in MAS [JIB07].

Motivated by the need to mitigate the adverse effect of unfair testimonies, the ultimate goal of this thesis is to propose a generic method which can be applied to most computational trust models to make them robust in the presence of unfair testimonies.

1.4 Contributions

By accomplishing the goal set out in the previous section, the research presented in this thesis has made the following contributions:

- (i) Although recent years have seen several attempts to mitigate the adverse effect of unfair testimonies, the proposals are generally not readily applicable in realistic settings as they require additional mechanisms or knowledge which is not always available in real environments, e.g. those in [JF03, SM03]. This research proposes a testimony-filtering method to address the adverse effect of the unfair testimonies. The filtering method grounds itself on the traditional interpretation of trust as dealing with the uncertainty about trustee's

behavior, and identifies the possible unfair testimonies by measuring the uncertainty that each testimony contains. Compared with the existing work, this filtering method does not require any additional information and mechanism other than the testimonies exchanged among agents. Moreover, experimental results show that it outperforms related work, both in terms of the processing time and the effectiveness in mitigating the adverse effect of the unfair testimonies.

- (ii) The filtering method is then modified to incorporate a novel *credibility* model. The credibility model equips the truster with the ability to evaluate an agent's (i.e. witness') credibility. In this thesis, a witness' credibility is essentially the evaluation of the witness' testimonies' usefulness in reducing the truster's uncertainty about the trustee's behavior. The credibility metric is derived according to the effect of third-party testimonies in the traditional research of trust, which is to make the truster "more certain of its trust" on the trustees [BK95].

Then the testimonies are filtered and aggregated according to the corresponding witnesses' credibility to mitigate the adverse effect of the unfair testimonies. A conservative view towards the testimonies is employed in aggregating the testimonies to further mitigate the adverse effect of unfair testimonies, which considers each testimony as a possible unfair testimony in the first place and adjusts it thereafter based on the corresponding witness' past testimonies.

In this thesis, *trust* and *credibility* are two different, though related, concepts. The truster's evaluation of *trust* in the trustee captures its uncertainty about the trustee's behavior, while its evaluation of an agent's (i.e. witness') *credibility* is the evaluation of the witness' testimonies' usefulness in reducing the truster's uncertainty. Correspondingly, the credibility model is decoupled from the trust model. Thus, it can be applied to most trust models, which makes it a generic solution to the problem caused by the presence of unfair testimonies.

- (iii) Unlike the existing work, which usually assumes that the agents would directly ask all the other agents in a MAS for testimonies, this thesis discards this assumption and proposes a

credibility-aware referral process on top of the proposed credibility model. The credibility-aware referral process facilitates agents' testimony discovery in an efficient manner, in which more credible agents are iteratively requested to discover testimonies by the testimony discovery initiator.

During the referral process, there may be some agents who are fair but recommend agents who give unfair testimonies. Such agents are termed as *malicious referrers*. The presence of the malicious referrers aggravates the adverse effect of the unfair testimonies. However, the presence of malicious referrers is generally left unaddressed in current research on trust modeling in MAS. This research also goes beyond merely addressing the presence of the unfair testimonies, and is the first to address this issue. An approach is proposed to counteract the malicious referrers based on the theory of *spreading activation* [CL75].

The proposed research is applicable to the general context of multi-agent systems (MAS). That is, it can be applied to systems that can be modeled as multi-agent systems (MAS), e.g. e-commerce systems, the Peer-to-Peer content-sharing networks, the Grid and so on.

1.5 Outline of the Thesis

This thesis is organized in the following manner.

- Chapter 2 reviews the notable work in the research on trust in MAS, and discusses the methods they employed to address the presence of unfair testimonies.
- Chapter 3 introduces trust into MAS by exploiting a widely used basic trust model, and then proposes a filtering method to mitigate the adverse effect of the unfair testimonies.
- Chapter 4 revamps the design of the filtering method with a novel credibility model.
- Chapter 5 empirically evaluates the effectiveness of the credibility model.
- Chapter 6 presents the credibility-aware referral process, which facilitates the agents to find relevant witnesses. Moreover, this chapter presents an approach to address the existence

of malicious referrers, which is left unaddressed in current research on trust modeling in MAS.

- Finally, Chapter 7 outlines the possible directions for future work, and concludes the thesis.

Chapter 2

Literature Review

After Marsh’s seminal work in [Mar94], which formally introduced *trust* as a measurable property in the domain of computer science, there has been a significant body of work in the research on trust modeling in the MAS. Many computational trust models have been proposed, which equip the agents with the ability to quantitatively evaluate the trust they should place in their interaction partners.

This chapter gives a review of a number of notable works in the research on trust modeling in MAS and studies how they address the presence of unfair testimonies.

2.1 Factors of a Computational Trust Model

The term “computational trust model” implies that the trust models adopt a *probabilistic* view of trust. That is, as discussed in Section 1.1, trust is considered as an agent’s subjective evaluation of the probability that its interaction partner would take a particular action in future interactions. There is another view of trust, namely, *cognitive* view, which models trust as made up of underlying beliefs, and where trust is a function of the degrees of those beliefs. To the best of our knowledge, there is currently only one trust model that adopts such a cognitive view of trust, which is proposed by Castelfranchi and Falcone [CF98]. A major disadvantage of this model is that it only gives an analysis of the underlying components of trust, but does not provide a feasible and practical approach to quantify trust based on those components. Hence, it is not suitable for MAS in general, and is not included in this review.

Besides the view of trust that is adopted, there are other factors for building a computational trust model. Before the review, we first identify a set of such factors.

2.1.1 Paradigm of the Model

The first factor that needs consideration is the paradigm of the model, namely, whether the model is centralized or decentralized. Most of the existing models are decentralized to cope with the decentralized and open nature of MAS. There is also a small number of models that are centralized.

If a decentralized trust model is deployed in MAS, each individual agent maintains past interaction experience with the other agents in its own memory. It also evaluates the trustworthiness of the other agents without depending on the computation of a central entity. The trust evaluation is always conducted by one agent on another agent. The two agents involved in the trust evaluation are called *truster* and *trustee* respectively. The one who is evaluating the trustworthiness of another agent is the *truster*, while the one whose trustworthiness is being evaluated by the *truster* is the *trustee*. Since each truster evaluates the trustee based on its own information regarding the trustee's past behavior, *trust* is considered as a subjective property of the trustee, and different trusters can have different perceptions of a trustee.

If a centralized trust model is deployed, agents do not maintain past interaction experiences with the other agents in their own memories. Instead, a central repository is set up to collect and store agents' past interaction experiences. There is also a central entity that evaluates the trustworthiness of other agents based on the experiences stored in the central repository. Every time an agent needs to evaluate trust on its interaction partner, it will turn to the central entity, who evaluates the trustworthiness of the trustee and reports the outcome to the truster (i.e., the agent who contacts the central entity for trust evaluation). Different trusters share the same trustworthiness of a trustee. This information is updated each time any agent updates the central repository with new interaction experience with the trustee. The centralized trust models are usually applied in scenarios where participating agents in MAS have a common point for interaction, such as the online transaction website like eBay.com, or where the agents' capabilities are too limited to carry out the trust evaluations by themselves.

2.1.2 Sources of Trust Information

As stated in Chapter 1, there are two primary sources of trust information that agents can use to evaluate trust on their interaction partners, namely, the personal interaction experience and the third-party testimonies. It is noted that such classification of sources of trust information is only applicable to decentralized models. In centralized models, the central entity will carry out the trust evaluations based on the available information maintained by the central repository. The available information for the central entity is always testimonies reported by all the participating agents in MAS.

The personal interaction experience is the most direct source of information for evaluating trust. This is because it is maintained by the trusters themselves and a rational agent would not hide any interaction experience from itself and cheat itself. There are also cases when truster agents' personal interaction experiences are not sufficient to evaluate trust in their interaction partners due to the lack of interactions. For example, an agent may have rarely or never interacted with another participating agent. In this case, agents would like to hear from the testimonies from a third-party regarding their interaction experiences with the same trustee. However, aggregating third-party testimonies induces the presence of unfair testimonies, which usually biases the trusters' trust evaluations and deteriorates the performance of the trust model. Some trust models do not incorporate this source of information to avoid the adverse effect caused by unfair testimonies.

Other than these two commonly-used sources of information, the social relationships between the agents (e.g. from same organization, or a friendship relationship) may also be taken into account in a number of models, e.g. [SM03, HJS06b]. The number of the models that use this kind of information is limited since such information is not always available in realistic settings. Inclusion of such information increases the models' dependence on additional mechanisms.

2.1.3 Granularity of Trust Information

After choosing the trust information, another factor that needs to be considered is the granularity of trust information. Trust information is essentially the feedback of the agents' behavior in the

past interactions, and is usually represented in the form of ratings. The ratings can be coarse-grained or fine-grained.

Ratings, can be coarse-grained, e.g. binary ratings. With the application of binary ratings, the behavior of an agent in an interaction is either “good” (denoted as 1) or “bad” (denoted as 0). The advantage of binary ratings is that it supports the use of the beta-family of probability density functions (PDFs), which provides a solid mathematical basis for evaluating trust.

Some models use finer-grained ratings, such as ordinal (or nominal) ratings with a scale larger than 2. Finer-grained ratings allows a better discrimination of the agents’ behaviors. In general, the ability of the ratings to discriminate agents’ behaviors is enhanced with the extension of the rating scale. In some models, the ratings are also associated with text labels, such as *very trustworthy*, *trustworthy*, *untrustworthy*, and *very untrustworthy*.

There are also models that apply continuous ratings drawn from a bounded range. Normally, the range would be $[0, 1]$ to capture the probabilistic nature of trust. Continuous ratings have the finest granularity, and provide the richest expressive power in discriminating agents’ behavior. Nevertheless, usage of continuous ratings requires extra overhead to maintain the trust information. In the models that use discrete ratings, the trust information can be maintained efficiently by only storing the counts of corresponding ratings instead of keeping individual rating after each interaction. In contrast, each rating must be kept if the ratings are continuous, which causes extra storage overhead for the agents.

It is noted that the granularity of trust information is not directly related to the granularity of the outcome of trust evaluation (i.e. the trustworthiness of the trustee). For example, the outcomes of the trust evaluations in the *Beta Reputation System* take continuous values, while the trust information used to make the trust evaluations is binary.

Other than the numeric ratings, there are also cases that free-text based user comments are used. They can be viewed as trust information with the highest granularity. However, such information would not be available in all scenarios. There are many scenarios that such information not hard to obtain and maintain, e.g. peer-to-peer content-sharing networks. Moreover, the free-text based comments, if available, are usually present together with ratings of certain granularity

as discussed above. They are usually optional and complimentary to the ratings. For example, in eBay.com, the buyer can choose to leave comment, but the rating is always required from the buyer when he give feedback for the seller. Moreover, to understand those user comments usually depends on NLP (natural language processing) techniques, which is beyond the scope of this thesis. Therefore, this thesis only puts focus on the rating-based trust information.

2.1.4 Context-dependence

The models can either be *context-dependent* or *context-independent*. A *context-dependent* model equips the agents with the ability to deal with more than one context and to maintain different trustworthiness for each trustee in different contexts considered¹. A *context-independent* model, on the other hand, only enables agents to maintain a single trustworthiness for each trustee without the discrimination of contexts.

As discussed in Section 1.1, trust, by its nature, is context-dependent. Hence, ideally, trust models should enable agents to maintain different trustworthiness for their interaction partners in different contexts considered. However, the ability to deal with multiple contexts rests on the availability of trust information in different contexts. This introduces the overhead of maintaining information in more than one context for the agents. Moreover, the ability to deal with multiple contexts normally increases the complexity of the models. Therefore, there is usually a trade-off between the ability to support multiple contexts and the extra maintenance overhead and increase of complexity.

In the literature, the number of existing trust models that deal with multiple contexts is limited. Most existing models only handle single context. This is reasonable and practical because most MASs focus on specific domains with agents performing limited and similar tasks. This makes it possible to generalize agents' behaviors into one single context. For example, consider there is a hierarchy of contexts, "service" is the parent context, while "storage service" and "communication service" are two child classes of "service". "Storage service" and "communication

¹Here, *context* refers to the settings in which agents' interactions are carried out. For example, there can be a context of "storage service", in which agents interact with each other for storage service. Similarly, there can be another context of "communication service".

service” are two different contexts, since an agent’s behavior in providing storage service does not necessarily imply anything about its behavior in providing communication service. However, if we do not differentiate the service an agent provides, we can consider its behaviors in “storage service” and “communication service” uniformly as “service”. Thus, agents’ behaviors in two contexts can then be generalized in one single context as “service”. Such a generalization has its advantage: agents’ interactions with others in different contexts are now fused into one single context. The result is an accumulation of more trust information, and hence a reduction in uncertainty in the trustees’ behavior.

2.1.5 Methods to Tackle Unfair Testimonies

As discussed earlier, the presence of unfair testimonies has an adverse effect on the performance of the computational trust models. A computational trust model is expected to be able to mitigate the adverse influence of the unfair testimonies. It is noted that this factor is only applicable to models that aggregate others’ testimonies.

Although the importance of mitigating the adverse effect of unfair testimonies is well recognized [JIB07], at the early stage of the research, trust models usually assume that all the agents would tell the truth when sharing the testimonies and ignore the presence of unfair testimonies, for example, work by Marsh [Mar94]. Recent years have seen several attempts to address this issue. According to Jøsang et al.’s classification in [JIB07], the methods to address the presence of unfair testimonies can be classified into two general categories: either *endogenous* methods or *exogenous* methods. An *endogenous* method is one that identifies the unfair testimonies by analyzing and comparing the testimonies themselves. In general, *endogenous* methods implicitly assume that unfair testimonies would be a minority amongst all the available testimonies. Based on this assumption, a testimony is considered as a possible unfair testimony if it deviates in some way from what the majority says. On the other hand, *exogenous* methods operate independently of the majority opinion, and use externally determined “fairness” of a testimony to choose the strategy in dealing with that testimony.

Having identified the factors that a computational trust model should consider, the next section presents a review of notable trust models in literature. The reviewed models are first classified into two main categories based on the *paradigm of model*, and then further classified into sub-categories depending on the factor of *method to tackle the unfair testimonies* (mainly for the decentralized models). A taxonomy of the reviewed models along the aforementioned two factors is given below:

Figure 2.1: Taxonomy of the review trust models

2.2 Centralized Computational Trust Models

In this section, we review a number of centralized computational trust models.

2.2.1 eBay Trust Model

A typical example of centralized model is the one used in `eBay.com`. In `eBay.com`, users give ratings to report their transaction partners' behaviors, which are stored and maintained by a centralized repository. There are 3 possible ratings, namely, `negative`, `neutral`, and `positive`. The ratings are given without discrimination of contexts. That is, ratings are given as feedback of a seller's general behavior without the discrimination of the items sold. Hence, the trustworthiness of a trustee is context-independent. The trustworthiness of a trustee is derived as the sum of the ratings that it has accumulated over the past 1 month, 6 months, and 12 months respectively. The trustworthiness of users is shared among all the members in the `eBay`

community, so that all the members in the community can know about their potential partners' behaviors.

It is not uncommon that some users in the *eBay* community would give unfair ratings for or against certain users. Since the trustworthiness is obtained by summation of all the ratings a trustee has accumulated, this model is susceptible to the presence of unfair ratings. However, this model does not have explicit mechanisms to address the presence of unfair ratings.

2.2.2 Jurca and Faltings

In the model proposed by Jurca and Faltings [JF03, JF04], the trustworthiness of agents are evaluated by *R-agent*. *R-agent* is a special type of agent in MAS, and acts as the central entity, whose task is to collect and aggregate ratings about normal agents' behavior to evaluate the trustworthiness. The ratings used to generate the trustworthiness are binary. Trustworthiness of an agent is generated as the average of all the ratings (regardless of the contexts) that this agent has accumulated over time. After the evaluations, the results will be sold back to normal agents when requested. Although the trustworthiness of the agents are evaluated by *R-agents* centrally, there is more than one *R-agent* distributed in MAS. Therefore, this model can be considered as a mixed model, but not a pure centralized model.

This model employs a preventive mechanism to tackle the presence of unfair ratings. It proposes a payment mechanism to incentivize agents to tell the truth when reporting ratings. This mechanism rewards agents who tell the truth when reporting ratings, and penalizes those who give unfair testimonies. By doing so, agents who tell the truth will gradually gain money, while those who give unfair testimonies will gradually lose money. Hence, this payment mechanism prevents the presence of unfair testimonies by making it rational for agents to share fair testimonies. However, in order to maintain this property, the payment mechanism is required to be separated from the one used to regulate agents' ordinary interactions. Nevertheless, in realistic settings, it is not practical to establish two separate payment systems that can be accepted by agents representing different stakeholders.

2.3 Decentralized Computational Trust Models

As the centralized model manages trust for all the agents in a centralized manner, it somewhat contradicts the decentralized nature of MAS. Centralized models do not support subjective trust evaluations. That is, different trusters see the same trustworthiness of a trustee. This makes it hard for the models to keep up with the trend that multi-agent systems are moving towards the open architecture. Such centralized models may be suitable for scenarios in which agents have same point of interaction, or those in which agents have limited capabilities. However, the central authority will easily be overloaded and become the single-point of failure when the number of participating agents grows.

Agents' capabilities, both in terms of memory, and reasoning and decision-making ability, have improved significantly in recent years. More and more computational trust models are designed to equip the agents with the ability to carry out trust evaluations individually. This section reviews a number of notable decentralized models, and the reviewed models are further classified based on the *methods to tackle unfair testimonies*.

2.3.1 Models without Method to Tackle the Unfair Testimonies

In general, models that are unable to tackle unfair testimonies are generally impractical since their performance will be adversely affected by the presence of unfair testimonies. Nevertheless, there is one noteworthy model that utilizes trust information other than the two commonly-used sources, i.e. personal interaction experiences and third-party testimonies.

2.3.1.1 FIRE

The FIRE model proposed by Huynh et al. [HJS06b] is a context-dependent model. In this model, the overall trust on a trustee is made up of four components: *interaction trust*, *witness reputation*, *role-based trust*, and *certified reputation*. Different sources of trust information are used to derive different components of trust. The first two components of trust are based on the personal interaction experiences and third-party testimonies.

Role-based trust is deduced from a set of predefined role-based rules. Role-based rules are domain-specific rules which define trust based on the social relationships between the two agents. For example, an agent may be preset to trust any other agent that is owned by the same user, or to trust another agent if it is a member of a well-known organization. Role-based trust in FIRE can be considered as a combination of the *system reputation* and *neighborhood reputation* in the ReGreT².

The main contribution of FIRE is that it introduces the concept of *Certified Reputation* [HJS06a], which is obtained from the third-party references provided by the trustee. Traditionally, the truster needs to actively collect third-party testimonies on the trustee. With the introduction of third-party references provided by the trustee, the burden of obtaining trust information is moved from the truster to the trustee.

Except for the role-based trust, which deduces trust based on predetermined rules, the other three components determine trust from information about the trustee agent's past behaviors. All three components obtain trust as a weighted mean of all the available ratings with the rating recency as weight. The recency of a rating is obtained based on the time difference between the current time and the time when the rating is collected. After each component finishes the evaluation, an overall trust value is obtained as a weighted mean of all the trust evaluations from the four individual components. The weights assigned to each component is predefined and fixed.

The introduction of the certified reputation increases this model's availability in various settings. It overcomes the difficulties that the other models usually encounter, such as a new agent's lack of interaction experiences with the existing agents and the difficulty in finding witnesses to share testimonies. However, this model depends on the social relationship to evaluate the role-based trust. This increases the dependence on the extra mechanism to make such information available. Furthermore, this model assumes that all the agents are honest when sharing testimonies or giving references. Mechanisms are needed to protect it from adverse influence of unfair testimonies (for the component of witness reputation) or references (for the component of certified reputation).

²ReGreT model is discussed in later section

2.3.2 Models with Endogenous Method to Tackle the Unfair Testimonies

2.3.2.1 Beta Reputation System

A typical decentralized model with an endogenous method to tackle the presence of unfair testimonies is the Beta Reputation System³, which is proposed by Jøsang et al. [JI02,WJI04]. It is a context-independent model.

With this model, an agent gives a *positive* rating of 1 to its interaction partner if the partner behaves cooperatively and it is satisfied with the partner's behavior in the interaction. Otherwise, a *negative* rating of 0 is given. The interaction experience with a trustee is basically a summary of the past ratings on that trustee, which is obtained as the counts of positive rating and negative rating ($\#positive$ and $\#negative$) in past interactions with that trustee. Sometimes older ratings are gradually forgotten with a decaying factor. Thus, the newer ratings exert a larger impact in calculating $\#positive$ and $\#negative$ than older ratings.

The behavior encountered by each individual agent is a binary event, which is modeled by a Beta distribution. Beta distribution is usually used to represent the posterior probabilities of binary events [CB90]. This provides a sound mathematical ground for agents to evaluate trust. The beta-family of probability density function (PDF) is a set of continuous functions indexed by two parameters α and β . In this model, α is set to be $\#positive + 1$, and β is set to be $\#negative + 1$. As discussed in Chapter 1, trust is basically the expectation that the agent would behave cooperatively in future. Therefore, a truster's evaluation of the trustee's trustworthiness is quantified as the expectation value of the Beta distribution, which is given by $\frac{\alpha}{\alpha+\beta} = \frac{\#positive+1}{\#positive+\#negative+2}$.

Other than depending on its private interaction experiences, the truster also aggregates third-party's testimonies when it is evaluating the trustee's trustworthiness. A testimony given by an agent is in the form of the counts of positive and negative ratings in this agent's previous interactions with the trustee. The aggregation is done by summing all the counts of positive and negative ratings in the testimonies and the counts of positive and negative ratings in the truster's personal interaction experiences. The trustee's trustworthiness is derived based on the aggregated $\#positive$ and $\#negative$.

³The original design of Beta Reputation System does not have a mechanism to address the unfair testimonies. It is then improved with an endogenous method.

Since the aggregation is simply a linear combination of the truster's private interaction experience and third-party testimonies, the trust evaluations are easily manipulated by unfair testimonies. However, the original design of Beta Reputation System does not show how it is able to tackle the presence of unfair testimonies. Although the original design lacks a mechanism to tackle the presence of unfair testimonies, it is the foundation of many computational trust models, such as TRAVOS [TPJL06] and the work in [Mui03].

Whitby et al. improve the original design of Beta Reputation System with an endogenous method to tackle the presence of unfair testimonies [WJI04]. This method first generates a majority opinion by aggregating all the available testimonies. The majority opinion, as an aggregation of individual testimony, also represents a Beta distribution. It identifies the possible unfair testimony by testing whether a testimony is outside the q quantile and $(1 - q)$ quantile of the majority opinion. If the test outcome is positive, the testimony is considered as a possible unfair testimony and will be discarded. Then the majority opinion is generated again with the remaining testimonies before a new round of unfair testimonies filtering starts. This process continues until no testimony can be discarded.

This filtering method implicitly assumes that unfair testimonies would be a minority among all the available testimonies. Consequently, it becomes inefficient when the majority of the testimonies are unfair. Moreover, this method is time-consuming because of its iterative nature.

2.3.3 Models with Exogenous Method to Tackle the Unfair Testimonies

2.3.3.1 Abdul-Rahman and Hailes

Abdul-Rahman and Hailes's model [ARH00] takes into account both the personal interaction experiences and third-party testimonies. This model is context-dependent. The trust information is given in the form of ratings with labels: *Very Trustworthy*, *Trustworthy*, *Untrustworthy*, and *Very Untrustworthy*.

Although this model takes into account both the personal interaction experience and third-party testimonies, the trust evaluations are based on the testimonies only. The personal interaction experiences are used only to determine the witness agents' *recommender trust*. Truster agent x 's

evaluation of a witness agent i 's recommender trust is determined based on the historical semantic distances between agent i 's testimonies on the trustees and agent x 's personal experiences with the same trustees. Agent i 's recommender trust would be high if the historical semantic distances are small, and vice versa.

This model acknowledges the existence of unfair testimonies. The influence of unfair testimonies is mitigated by assigning different weights to testimonies given by different agents. The weights are determined based on their corresponding recommender trust. Moreover, this model is one of the first attempts to adjust testimony before aggregation. The adverse influence of unfair testimonies is further reduced by adjusting the testimonies from different agents based on historical semantic distances. For example, truster agent x receives a testimony from agent i , which states that the trustee agent y is *Very Trustworthy*. But agent x finds out later that after the interaction with agent y its own trust evaluation is only *Trustworthy*. In this case, the semantic distance between x and i is defined as -1 . Subsequently, when agent i gives another testimony to agent x , agent x will adjust the testimony accordingly before aggregating the testimony.

The problem of the method to address the presence of unfair testimonies is that it predefines the weights assigned to witnesses with various levels of recommender trust in an ad-hoc manner. Furthermore, the adjustment of testimony only takes into account the semantic distance in the last recommendation. Such a treatment is not sufficient to represent the true difference between the views of the truster and the witness.

2.3.3.2 ReGreT

ReGreT [SM03] is a context-dependent model (context is called *aspect* in this model). It takes into account both the personal interaction experiences (termed as *impressions* in this model) and third-party testimonies. Other than these two sources of information, it also considers information derived from the social structure among agents.

Agent x 's *direct trust* on agent y is obtained based on x 's previous impressions on y . Each impression is agent x 's subjective evaluation of the outcome of an interaction in a certain aspect. The outcome of the evaluation is in the form of continuous rating in the range of $[-1, 1]$. The

direct trust on agent y in a specific aspect is calculated as the weighted mean of all the ratings associated with that aspect. Each rating is weighted according to its recency, i.e. a more recent rating is assigned a higher weight, and vice versa. ReGreT also measures the reliability of each direct trust evaluation. The reliability measurement is obtained based on two factors: the number of ratings taken into account when generating the evaluation and the deviation of the ratings. If the agent has a reliable direct trust evaluation, it will use it as the final evaluation of trust. Otherwise it will also use reputations. There are four types of reputation taken into account, namely, *witness reputation*, *neighborhood reputation*, *system reputation*, and *default reputation*.

Witness reputation is obtained by aggregating third-party testimonies. To obtain *neighborhood reputation*, the truster first obtains its direct trust on the agents in the trustee's neighborhood. Then a set of predefined fuzzy rules is used to generate the neighborhood reputation based on the direct trust on the trustee's neighbors and their relationship with the trustee. The *system reputation* is predetermined according to the trustee's social roles. If there is totally no information available about the trustee, *default reputation* is used as the last resort. Default reputation is a predefined constant.

ReGreT acknowledges the existence of unfair testimonies when deriving the witness reputation. The adverse effect of unfair testimonies is reduced by generating the witness reputation as the weighted mean of all the available testimonies. The weight is determined based on the social structure among the witness, the truster, and the trustee, and a set of predefined fuzzy rules are used to determine the weights for the witness agents.

The ReGreT model depends on the social structure to obtain the witness reputation (and neighborhood reputation and system reputation as well). However, such information is not always available in MAS. To make such social information available, extra mechanism is needed. For example, each agent needs to explicitly maintain a social network. This increases the model's dependence on additional mechanisms. Furthermore, the description of the ReGreT model in [SM03] does not specify how such information is made available to the agent. Thus, the method to address the presence of unfair testimonies is not readily applicable in realistic setting.

2.3.3.3 Yu And Singh

The model proposed by Yu and Singh [YS02] takes into account both private interaction experiences and third-party testimonies. It is a context-dependent model.

In this model, private interaction experiences and testimonies are sets of ratings that reflect the quality of the past interactions. Only interactions within a fixed temporal window are considered in the trust evaluations. Each agent x defines a lower bound ω_i and an upper bound Ω_i . With ω_i and Ω_i , agent x separates the whole distribution of the available ratings into three regions, namely, (1) the region of ratings that contributes to agent x 's evaluation that the trustee y is untrustworthy, (2) the region that contributes to the evaluation that the trustee has no clear classification between untrustworthy and trustworthy, and (3) the region that contributes to the evaluation that the trustee is trustworthy. Then the ratings in different regions are counted to derive a *basic probability assignment (bpa)* over the distribution of the ratings. Finally, trustworthiness is evaluated based on the *bpa* using the Dempster-Shafer theory of evidence [Sha76, Dem68]. Trustworthiness in this model is in fact a set of three values, which correspond to the probabilities that the quality of future interaction with the trustee will fall in the aforementioned three regions.

Most of the notable work assumes that an agent would ask for testimonies from all the other agents in MAS. Such an assumption is hard to hold since there is usually a large number of agents within a MAS. The communication cost would be prohibitively high if each individual agent asks for testimonies from every other agent. This model is the only existing work that proposes an approach for agents to collect testimonies efficiently. When agent x is evaluating the trustworthiness of agent y , it sends out queries to its neighbors. A neighbor is an agent that agent x has direct connection with. One of its neighbors, upon receiving the query, will return testimony to x if it did interact with y before. Otherwise, it will return *referrals*. Similarly, those referrals, when queried by x , can provide testimonies or generate referrals iteratively.

The presence of unfair testimonies is addressed by introducing a variant of the Weighted Majority Algorithm (WMA) [LW94] called WMA Continuous (WMC) [YS03]. With WMC, each agent maintains a weight for each of the witnesses whose testimonies have been aggregated before. The weight basically captures how useful the witness agent is in giving testimonies. In

order to apply the WMC, the testimonies (in the form of *bpa* over ratings) are converted to a scalar rating first. The weight is initialized as 1. Then it is updated each time the witness agent's testimony (now in the form of rating) is aggregated by the truster agent. If the testimony given by the witness (denoted as a_w) is similar to the truster's (denoted as a_s) evaluation of trust (it is also converted to scalar rating) in the trustee after interacting with that trustee, the witness' weight remains unchanged. Otherwise, as a penalty, the witness' weight (denoted as $TT_{s,w}$) is decreased with the following formula:

$$TT'_{s,w} = TT_{s,w} \cdot \theta \quad (2.1)$$

where $TT'_{s,w}$ is the witness' weight after update, θ denotes the amount of the penalty. It is determined according to the distance between agent a_w 's testimony (denoted as $TP_{w,o}$) and agent a_s 's post-interaction direct trust (denoted as $TP_{s,o}$) in the same trustee, and it can be any value that satisfies:

$$\delta^{|TP_{w,o} - TP_{s,o}|} \leq \theta \leq 1 - (1 - \delta)|TP_{w,o} - TP_{s,o}| \quad (2.2)$$

In [YS03], the upper bound $1 - (1 - \delta)|TP_{w,o} - TP_{s,o}|$ is chosen.

Then, with each witness agent's weight evaluated, third-party testimonies are aggregated as a weighted mean of all the collected testimonies.

Although this work does not depend on extra mechanism and knowledge, a primary problem is that WMC is too sensitive to unfair testimonies. That is, every testimony that is not *exactly* the same as the truster's evaluation of trust on the trustee will be considered as unfair, and the witness' weight will be decreased accordingly. Such a sensitive strategy would reduce the fair testimonies' contribution to the trust evaluations. This makes it even less effective than the strategy of not applying WMC in the case of few agents giving unfair testimonies.

2.3.3.4 Kafali and Yolum

Kafali and Yolum's work does not propose a complete trust model. Instead, it focuses on the factor of *methods to tackle unfair testimonies*, and proposes an exogenous method to combat unfair testimonies in the context of ART testbed⁴ [OKY06].

⁴ART testbed: <http://www.lips.utexas.edu/art-testbed/>

ART testbed simulates the interactions among agents in an art appraisal domain, in which agents function as painting appraisers with varying levels of expertise in different artistic eras. In the context of ART testbed, the trustee is in fact the painting, which is not a living object and does not have its own behavior. The trustworthiness of the trustee is basically the true value of the painting, and it is set randomly by the testbed, instead of being determined based on its past behaviors. The appraising agent (i.e. the truster) is assigned a number of paintings whose values are to be appraised. If the appraising agent does not have enough expertise to complete the appraisal, it can in turn buy opinions (i.e. third-party testimonies) from other appraiser agents. An appraiser agent may provide unfair testimonies regarding the values of paintings. Hence, before the appraising agent requests for opinions, it evaluates the reputation of other appraiser agents, and chooses those more reputable to request opinions from in order to tackle the unfair testimonies. An agent's *reputation* basically captures the usefulness of this agent's opinions [FKM⁺05, FKM⁺06, OKY06]. To evaluate an agent's reputation, the appraising agent would also buy others' reports on this agent's reputation. Similarly, the appraising agent evaluates the *sociability* of other appraisers before requesting for reputation reports. An agent's *sociability* denotes the usefulness of this appraiser's reputation reports [FKM⁺05, FKM⁺06, OKY06]. The appraising agent's final appraisal of the painting, i.e. the truster's evaluation of the trustee's trustworthiness, is derived as a weighted mean of all the opinions bought from other appraisers, with corresponding opinion provider's reputation as weight.

Kafali and Yolum's work proposes a method to combat unfair testimonies (both in the opinions about paintings and reputation reports). This method is similar to Yu and Singh's proposal [YS03], which adjusts an agent's reputation and sociability based on the distance between the true value of the painting and the agent's opinion. Upon knowing the true values of the paintings (i.e. the trustworthiness of the trustee), the appraising agent (i.e. the truster) updates its evaluation of other appraiser agents' reputation and sociability, which is conducted as follows:

$$\begin{aligned}
 Agent.Rep &= Agent.Rep * (1 + ((1 - Agent.Rep) * UpdateMargin)) \\
 RefAgent.Soc &= RefAgent.Soc * (1 + ((1 - RefAgent.Soc) * UpdateMargin)) \quad (2.3)
 \end{aligned}$$

where *UpdateMargin* is positive if the difference between the true value and the appraising agent's final appraisal of the painting (i.e. the appraisal error) is smaller than a threshold, and it is negative otherwise. In the above equations, *Agent* denotes an opinion provider (i.e. witness) who gives opinion on the painting to the appraising agent (i.e. truster), and *RefAgent* denotes a reputation provider who is previously consulted by the appraising agent to get reputation about *Agent*. *Agent.Rep* represents *Agent*'s reputation, while *RefAgent.Soc* denotes *RefAgent*'s sociability. Eq. (2.3) makes sure an agent's reputation and related referring agent's sociability are increased gradually if its opinion is close to the painting's true value.

Eq. (2.3) is applicable in the situation when the appraisal error is smaller than a threshold (i.e. *UpdateMargin* is positive). When the appraisal error is larger than the threshold (i.e. *UpdateMargin* is negative), *Agent*'s reputation and corresponding *RefAgent*'s sociability are updated as follows:

$$\begin{aligned} Agent.Rep &= Agent.Rep * (1 + UpdateMargin) \\ RefAgent.Soc &= RefAgent.Soc * (1 + UpdataMargin) \end{aligned} \quad (2.4)$$

With Eq. (2.4), an agent's reputation and corresponding referring agent's sociability are updated downward if it is possibly giving unfair opinions (i.e. the appraisal error is larger than the threshold). It is noted that in Eq. (2.4) an agent's reputation (or sociability) is updated downward faster than when it is increased (as in Eq. (2.3)). By doing so, it is able to quickly identify an agent who has continuously given unfair opinions (or reputation reports).

To combat the unfair testimonies, when the appraising agent requests opinions, it sorts all the available appraisers according to their reputation, and directs the requests to the top n appraisers from the sorted list.

This work does not depend on additional mechanism or knowledge to combat the unfair testimonies, and it is successful in the context of ART testbed. However, the description of its strategy in [OKY06] fails to make clear some key points of the strategy, which makes it an ad-hoc approach to the problem caused by the presence of unfair testimonies. For example,

- The rules to update agent's reputation and sociability are determined in an ad-hoc manner. It does not elaborate how these rules (Eq. (2.3) and (2.4)) are derived;
- *UpdateMargin* is important in the reputation and sociability update (Eq. (2.3) and (2.4)). Its value depends on the chosen threshold. However, it does not specify how this threshold is determined;
- The appraising agent chooses the top n appraisers with the highest reputation to request the opinions from. Nevertheless, the determination of n remains unclear in [OKY06].

Finally, we summarize the models reviewed in this chapter in Table 2.1.

2.4 Summary

The need to mitigate the adverse effect of unfair testimonies is a fundamental problem in trust modeling in MAS [JIB07]. However, there are still models that simply ignore the presence of unfair testimonies, such as Marsh's work [Mar94]. From the review carried out in this chapter, it is clear that the existing work is not readily applicable in realistic settings.

Against this background, this thesis aims to propose a generic approach that is applicable to most computational trust models to mitigate the adverse effects of unfair testimonies. The goal of this thesis is not to propose a new trust model. Instead, it focuses on method to tackle unfair testimonies. It assumes the following factors that were identified in Section 2.1.

- **The Paradigm**

Since the centralized models somewhat contradict the decentralized nature of MAS, this thesis does not consider the trust models with the centralized paradigm. The proposed research is applicable to the trust models with the decentralized paradigm.

- **Sources of Trust Information**

This research is applicable to the trust models that utilize personal interaction experiences and third-party testimonies. Third-party reference can be considered as a special type of third-party testimony. Thus, this research is also applicable to trust models with the inclusion of the third-party references.

Models	Paradigm	Source of Trust Information	Information Granularity	Context Dependence	Method to Tackle Unfair Testimonies
eBay Trust Model	Centralized	Personal Experiences + Testimonies	3-level discrete ratings	Context-Dependent	N/A
Jurca and Faltings [JF03, JF04]	Centralized	Personal Experiences + Testimonies	Binary ratings	Context-Dependent	Depending on additional payment mechanism ^a
FIRE [HJS06b]	Decentralized	Personal Experiences + Testimonies + Social Structure + Third-party References	Continuous in the range $[-1, 1]$	Context-Dependent	N/A
Beta Reputation System [JI02]	Decentralized	Personal Experiences + Testimonies	Binary ratings	Context-Independent	Endogenous ^b
Abdul-Rahman and Hailes [ARH00]	Decentralized	Personal Experiences + Testimonies	4-level discrete ratings	Context-Independent	Exogenous
Yu and Singh [YS02]	Decentralized	Personal Experiences + Testimonies	Continuous in the range $[0, 1]$	Context-Dependent	Exogenous
ReGreT [SM03]	Decentralized	Direct Experiences + Testimonies + Social Structure	Continuous in the range $[-1, 1]$	Context-Dependent	Exogenous and depending on additional social relationship information
Kafali and Yolum [OKY06]	Decentralized	N/A ^c	N/A ^c	N/A ^c	Exogenous

^aThis payment mechanism incentivizes agents to share fair testimonies.

^bThe original design does not address the presence of unfair testimonies. Whitby et al. have proposed an endogenous method to improve the original design in [WJI04]

^cThis work only focuses on the factor of *methods to tackle unfair testimonies*.

Table 2.1: Computational Trust Models

The social structure information is not always available in MAS. The inclusion of social structure information increases the applicability of the trust models at the cost of increasing the models' dependence on additional mechanisms to make such information available. Nevertheless, the inclusion of social structure information in trust models does not affect the applicability of the proposed approach, since it does not depend on social structure information. Hence, this research is also applicable to trust models with the inclusion of the social structure information provided that third-party testimonies are also included.

- **Granularity of Trust Information**

This research is applicable to the trust models with any granularity of the trust information, since the application of ratings with different granularity does not impact the agents' behavior in giving testimonies.

- **Context-dependence**

Trust, by its nature, is a context-dependent concept. Hence, a trust model should ideally be able to deal with multiple contexts to capture the context-dependent notion of the concept. However, in Table 2.1, there also exists a number of context-independent models which generalize agents' behaviors in multiple contexts into one single context. In fact, what makes a model context-dependent is its ability to make use of the trust information collected in multiple contexts and provide trust evaluations associated with different contexts. This is because trust models usually apply the same method to evaluate trustworthiness in different contexts based on the available information. In this respect, whether a model is context-dependent or context-independent does not have an impact over the applicability of the proposed approach. Therefore, the proposed approach is applicable to both context-dependent and context-independent trust models.

The following chapters document our effort to achieve the goal set out for this thesis.

Chapter 3

An Uncertainty-based Method To Filter Unfair Testimonies

In view that the presence of unfair testimonies would deteriorate the performance of the computational trust models, a fundamental problem in the research on trust modeling in MAS is to mitigate the adverse effect of the unfair testimonies. This chapter proposes an uncertainty-based method to address this problem, and also evaluates its effectiveness.

3.1 Introduction

Avoidance or mitigation of the adverse effect caused by the presence of unfair testimonies is a fundamental issue in trust modeling in MAS [JIB07]. Nevertheless, as reviewed in Chapter 2, this issue is left unaddressed in some trust models since they assume that agents would always tell the truth when reporting testimonies. Although there have been several attempts to address this issue in recent years, they are not readily applicable in realistic settings, e.g. those in [JF03, SM03]. Those models depend on extra mechanisms or knowledge, which are not always available in realistic settings.

This chapter proposes a filtering method to reduce the adverse effect of unfair testimonies. Apart from the testimonies exchanged among agents, it does not require any extra mechanism or knowledge. Recall that trust is basically about one individual's uncertainty about another individual's behavior in future interactions. In light of such an interpretation of trust, this filtering method measures the quality of each testimony based on the uncertainty it contains, and then

filters the testimony by checking its quality against the quality of the truster's current belief about the trustee.

Before presenting our proposal to filter the unfair testimonies, we introduce trust into MAS by exploiting an existing basic trust model, namely the Beta Reputation System (BRS) [JI02, JHF03]. This basic trust model is then extended with the proposed filtering method to mitigate the adverse effect of unfair testimonies. BRS is a typical model that uses binary ratings. A significant advantage of using BRS is that binary ratings support the use of the beta-family of probability density functions (PDFs). This provides a solid mathematical basis for evaluating trust.

As discussed in Section 2.4, to reduce the complexity of the basic trust model, this thesis makes an assumption that the trust model would generalize the agents' behaviors into one single context, though the proposed approach is applicable to both context-dependent and context-independent trust models.

This chapter is organized as follows. Section 3.2 introduces the basic trust model. This basic trust model is then extended with the filtering method to mitigate the adverse effect of unfair testimonies in Section 3.3. The filtering method is evaluated in Section 3.4. Finally, a discussion of the proposed filtering method is presented in Section 3.5 to conclude this chapter.

3.2 The Basic Trust Model

In this section, the main components of the basic trust model are described. For the ease of presentation, we first present some notations which will be used throughout this thesis.

3.2.1 Notations

We denote the set of all agents in MAS as $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$, where $a_i (i = 1, \dots, n)$ is an agent in the set. Before an agent interacts with another agent, it will evaluate the trust it should place in the latter. The one who is making the evaluation is the *truster*, while the one whose trustworthiness is being evaluated is the *trustee*. The truster is denoted as agent a_s , while the trustee is denoted as agent a_o . a_s 's evaluation of trust it should place in a_o is denoted as $TP_{s,o}$.

An agent who reports a testimony on trustee agent a_o to truster agent a_s is called a witness of agent a_o , and is denoted as agent a_w . $\mathcal{W}_{s,o}$ is used to denote the set of such witness agents that give testimonies on agent a_o to agent a_s .

As discussed earlier in Section 1.1, trust is basically about dealing with uncertainty of the trustee's behavior, which can be represented as the subjective evaluation of the probability that the trustee would behave cooperatively in future interactions. Hence, "predicting (the trustee's) future behavior" and "evaluating trust (on the trustee)" are used interchangeably sometimes in this thesis.

3.2.2 Maintaining Personal Interaction Experiences

For agent a_s , a_o 's behavior in an interaction can either be *satisfactory* or *unsatisfactory*. Whether an interaction is *satisfactory* or *unsatisfactory* is based on a_s 's own expectation of a_o 's behavior in the interaction. In general, a_s considers the interaction *satisfactory* if a_o behaves cooperatively and *unsatisfactory* otherwise. After an interaction with a_o is finished, a_s records in its memory a rating of *positive* for a_o if its behavior is satisfactory. Otherwise, a *negative* rating is recorded. Truster agent a_s 's rating for the trustee agent a_o in an interaction t can be represented as:

$$r_{s,o}^t = \begin{cases} r^+ & \text{if } a_o \text{'s behavior is satisfactory} \\ r^- & \text{otherwise} \end{cases} \quad (3.1)$$

Here, r^+ and r^- denote the positive and negative rating respectively, which can be represented as follows:

$$r^+ = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad r^- = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (3.2)$$

Instead of maintaining ratings of all the past interactions, each agent only maintains a summary of ratings for the past interactions. The summary of ratings is basically a summation of all the recorded ratings in its memory, which is calculated as:

$$R_{s,o}^{T_c} = \begin{bmatrix} P_{s,o}^{T_c} \\ N_{s,o}^{T_c} \end{bmatrix} = \sum_{t=1}^{T_c} r_{s,o}^t. \quad (3.3)$$

Here T_c is the time index of the interaction for which a_o 's trustworthiness is evaluated. t is the time index of the interaction after which a rating was recorded. $P_{s,o}^{T_c}$ and $N_{s,o}^{T_c}$ denote the counts of a_o 's satisfactory and unsatisfactory behavior respectively in a_s 's memory.

Agents may change their behaviors over time. Therefore, it is desirable that agents' past behaviors should be gradually forgotten. A decaying factor (λ) is usually applied to control the rate that agents' past behaviors are forgotten over time. Moreover, a truster usually maintains a summary of ratings for the past transactions within a window of size W . With these two factors applied, the summary rating can be obtained as:

$$R_{s,o}^{T_c} = \left[\frac{P_{s,o}^{T_c}}{N_{s,o}^{T_c}} \right] = \sum_{t=T_c-W+1}^{T_c} \lambda^{T_c-t} \cdot r_{s,o}^t, \quad \lambda < 1 \quad (3.4)$$

3.2.3 Making the Trust Evaluation

Agent a_o 's behavior in an interaction is a binary event, i.e. either satisfactory or unsatisfactory, and a_s 's evaluation of a_o 's trustworthiness is basically the subjective probability that a_o would behave satisfactorily (i.e. behave cooperatively) in future. The probability of such binary events can be represented by the beta family of probability density functions (PDF). The beta family of PDF is a set of continuous functions indexed by two parameters α and β [CB90]. α and β are two parameters controlling the shape of the beta distribution. Examples of beta distributions with various values of α and β are shown in Figure 3.1.

Figure 3.1: Examples of beta distribution

The probability density function (PDF) of the beta distribution is given by [CB90]:

$$f(X|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} X^{\alpha-1} (1-X)^{\beta-1};$$

$$0 < X < 1, \alpha > 0, \beta > 0 \quad (3.5)$$

where $B(\alpha, \beta)$ is the beta function. The expectation value of the beta distribution is calculated with the following formula:

$$E(X|\alpha, \beta) = \frac{\alpha}{\alpha + \beta} \quad (3.6)$$

Now consider how agent a_s can evaluate its trust on a_o based on the recorded ratings of agent a_o . In this case, the values of shape parameters α and β are basically the total counts of satisfactory and unsatisfactory interactions with agent a_o in agent a_s 's memory. Prior to any interaction with a_o , a_s 's uncertainty about a_o is maximized, since a_o has equal probabilities of behaving satisfactorily or unsatisfactorily. Hence, the initial setting for α and β are $\alpha = \beta = 1$, which makes $E(X|\alpha, \beta) = 0.5$. Then after an interaction with agent a_o , a_s obtains a new observation regarding a_o 's behavior. Based on standard techniques in Bayesian analysis, the values of parameters α and β are updated by adding the counts of agent a_o 's satisfactory and unsatisfactory behavior to the initial values of α and β respectively. According to [JI02, JHF03], that is done by:

$$\alpha = 1 + P_{s,o}^{T_c},$$

$$\beta = 1 + N_{s,o}^{T_c} \quad (3.7)$$

which is basically a Laplace-estimator [Ces90]. Then agent a_s 's trust on agent a_o , i.e. the probability that a_o would behave cooperatively, is calculated as the expectation value of the beta distribution as follows:

$$TP_{s,o} = \frac{\alpha}{\alpha + \beta} \approx \frac{1 + P_{s,o}^{T_c}}{2 + P_{s,o}^{T_c} + N_{s,o}^{T_c}} \quad (3.8)$$

3.2.4 Aggregating Third-party Testimonies

When a truster agent evaluates the trust that it should place in a trustee agent, it will also seek third-party testimonies on experience in interacting with the same trustee.

Each agent, when contacted by the truster agent, shares its own personal interaction experiences with the same trustee agent, which are also in the form of summary rating as in Eq. (3.4). In light of the new observations of the trustee agent's behaviors in the testimonies, the truster agent a_s updates parameters α and β of the beta distribution representing the trustee's behavior:

$$\begin{aligned}\alpha &= 1 + P_{s,o}^{T_c} + \sum_{w \in \mathcal{W}_{s,o}} P_{w,o}^{T_c}, \\ \beta &= 1 + N_{s,o}^{T_c} + \sum_{w \in \mathcal{W}_{s,o}} N_{w,o}^{T_c}\end{aligned}\tag{3.9}$$

where $\mathcal{W}_{s,o}$ is the set of witness agents, a_w is one of the witnesses in the set $\mathcal{W}_{s,o}$, and $P_{w,o}^{T_c}$, $N_{w,o}^{T_c}$ are the numbers of positive and negative ratings in a_w 's testimony. Then the truster agent updates its evaluation of $TP_{s,o}$ with the new value of α and β according to Eq. (3.8).

3.3 Extension with the Uncertainty-based Testimonies Filtering

Third-party testimonies are sought to reduce the truster's uncertainty regarding the trustee's behavior. However, if the third-party testimonies are aggregated blindly, the trust evaluations are easily biased by the presence of unfair testimonies. In the basic trust model presented in Section 3.2, agent a_s is now evaluating its trust on agent a_o , and asks for testimonies on a_o from the other agents. Some may give unfairly positive testimonies to agent a_s . Since the basic trust model aggregates testimonies by summing the counts of positive ratings and negative ratings directly, those unfairly positive testimonies will increase the value of α in Eq. (3.9). This would lead to the increase in a_s 's evaluation of trust on a_o according to Eq. (3.8).

In order to mitigate the adverse effect of unfair testimonies, we extend the basic trust model with a proposal to filter the unfair testimonies. The basic idea of the proposed filtering method is as follows. Since the fairness of a testimony is a subjective measure of the usefulness of the testimony, it is reasonable for a rational agent to aggregate testimonies that do not deviate (within an acceptable range) from what it currently believes about the trustee. Hence, before a new testimony is aggregated, the truster first measures the qualities of both its current belief about the

trustee and the yet-to-be-aggregated testimony on the trustee. Truster's current belief about the trustee is generated with its personal interaction experience and already-aggregated testimonies. If the new testimony shows a significant quality divergence from the quality of the current belief, the testimony deviates from what the truster perceives in some way. Thus it is considered as a possible unfair testimony and it will be filtered.

As discussed in Section 1.1, the truster's evaluation of trust on the trustee agent basically captures its uncertainty regarding the trustee agent's behavior. Therefore, the quality of the testimony (as well as the truster agent's current belief), based on which the trust evaluation is made, can be measured based on the "amount" of uncertainty it contains. Given this, *entropy*, a common measurement of information uncertainty [CT91], is employed as the basis to measure the testimony quality.

The entropy of a discrete random variable $\mathcal{V}$ is calculated as:

$$H(\mathcal{V}) = - \sum \Pr(v_i) \log(\Pr(v_i)) \quad (3.10)$$

where v_i is a possible value of $\mathcal{V}$, and $\Pr(v_i)$ is the probability of $\mathcal{V}$ taking the value v_i .

Suppose $\mathcal{V}$ can take n different values $v_1, v_2, \dots, v_n$. The maximum uncertainty H_{max} occurs when $\Pr(v_1) = \dots = \Pr(v_i) = \dots = \Pr(v_n)$, while the minimum uncertainty H_{min} occurs when $\mathcal{V}$ always takes only one possible value v_i [CT91].

Since binary ratings are employed in the basic trust model, uncertainty (about the trustee's behavior) contained in each testimony R can be measured as:

$$H(R) = - \Pr(\alpha) \log(\Pr(\alpha)) - \Pr(\beta) \log(\Pr(\beta)) \quad (3.11)$$

where $\Pr(\alpha)$ and $\Pr(\beta)$ are given by:

$$\Pr(\alpha) = \frac{\alpha}{\alpha + \beta}, \quad \Pr(\beta) = \frac{\beta}{\alpha + \beta}. \quad (3.12)$$

Here α and β are given by Eq. (3.7), which are basically one plus the positive counts and negative counts respectively in the received testimony. In the case of binary ratings, the maximum uncertainty $H_{max}(R)$ occurs when $\Pr(\alpha) = \Pr(\beta)$, i.e. testimony R is generated with equal

presence of satisfactory and unsatisfactory encounters in the past W interactions. The minimum uncertainty $H_{min}(R)$ occurs when testimony R is generated with only positive or negative ratings in the past W interactions (W is the window size in Eq. (3.4)).

Then, the *quality* of testimony R is measured according to the uncertainty it contains as follows:

$$\begin{aligned} Q(R) &= 1 - \frac{H(R) - H_{min}}{H_{max} - H_{min}} \\ &= \frac{H_{max} - H(R)}{H_{max} - H_{min}} \end{aligned} \quad (3.13)$$

The quality of the truster a_s 's personal interaction experience with the trustee a_o , denoted as $Q(R_{s,o})$, is measured similarly. When measuring $Q(R_{s,o})$, α and β become one plus the positive counts and negative counts respectively in the truster's own experience with the trustee.

The testimonies filtering and aggregation starts with evaluating the quality of the truster's current belief about the trustee. Before aggregating any testimony, the truster's current belief about the trustee is generated with only the truster a_s 's personal interaction experience with the trustee. In other words, the quality of the current belief $Q(R_M)$ is initialized as $Q(R_M) = Q(R_{s,o})$. After that, truster a_s considers each new testimony R as a possible unfair testimony if:

$$|Q(R) - Q(R_M)| > \varepsilon \quad (3.14)$$

where $Q(R)$ is the quality of the yet-to-be-aggregated new testimony R , and $Q(R_M)$ is the quality of truster agent a_s 's current belief about trustee a_o . ε is the filtering threshold ($\varepsilon \in [0, 1]$), which controls the sensitivity to the presence of unfair testimonies. With a larger ε , the truster is less sensitive to the unfair testimonies, whereas with a smaller ε , the truster is more sensitive.

If the testimony is considered to be a possible unfair testimony, it is discarded. Otherwise, it is aggregated by the truster a_s to update its current belief about trustee a_o , and the quality of the current belief ($Q(R_M)$) is updated accordingly. The same process is carried out on each new testimony. The complete procedure to filter and aggregate testimonies is shown in Figure 3.2.

Procedure *Testimonies_Filtering*(s, o) a_s : the truster agent, a_o : the trustee agent, a_w : the agent whose testimony is currently being processed by a_s .

- 1: a_s measures the quality of its personal interaction experience with a_o , i.e. $Q(R_{s,o})$ using Eq. (3.13)
- 2: a_s initializes the quality of its current belief about a_o as $Q(R_M) = Q(R_{s,o})$
- 3: **for all** testimonies **do**
- 4: a_s measures the quality of the testimony reported by a_w , i.e. $Q(R_{w,o})$ using Eq. (3.13)
- 5: **if** $|Q(R_{w,o}) - Q(R_M)| > \varepsilon$ **then**
- 6: a_s discards a_w 's testimony
- 7: **else**
- 8: a_s aggregates a_w 's testimony, and update α and β using Eq. (3.9)
- 9: a_s updates the quality ($Q(R_M)$) of its current belief about a_o accordingly
- 10: **end if**
- 11: **end for**
- 12: a_s evaluates trust on agent a_o with the current belief

Figure 3.2: Uncertainty-based testimonies filtering and aggregation

3.4 Experimental Study

A number of experiments has been conducted to evaluate the performance of the proposed testimony-filtering method. Since the validity and effectiveness of the basic trust model have already been studied in [JI02, JHF03], the focus of the experiments is to investigate the effectiveness of the extension, namely the uncertainty-based testimony-filtering method.

3.4.1 Experimental Methodology

The experiments are conducted by simulating interactions among 100 agents. Those agents are separated into two sets: there are 99 agents in *Set A* and the rest is in *Set B*. The agent in *Set B* provides services that the agents in *Set A* need to use. The agents in *Set A* are called the *service consumers*, while the agent in *Set B* is called the *service provider*. Each consumer in *Set A* carries out 1000 interactions with the same provider in *Set B*¹. After each interaction, the consumer gives a rating to its interaction partner, i.e. the service provider, and the rating is stored

¹This may not be the case in the realistic settings, in which there are usually more than one service provider, and an agent usually has few interactions with same interaction partner [RZ02]. However, this configuration facilitates an intensive study of the unfair testimonies and the proposed filtering method's effectiveness, which is the focus of the experiments

locally in its memory. The consumer gives a positive rating if the provider behaves cooperatively (e.g. provides a good service). Otherwise, it gives a negative rating.

Beta Reputation System is applied by the consumers to evaluate the trustworthiness of the providers before each interaction. The number of ratings that are taken into account when generating the rating summary is set as 10, i.e. $W = 10$ in Eq. (3.4). A decaying factor is applied, and it is set to 0.9, i.e. $\lambda = 0.9$ in Eq. (3.4). The threshold of the filtering is set to 0.3, i.e. $\varepsilon = 0.3$ in Eq. (3.14).

The experiments simulate the provider by controlling its behaviors with its *willingness*. The provider's willingness is essentially the probability that it would behave cooperatively with the consumer agent. It takes values in the range of $[0, 1]$. The higher the willingness, the more likely the provider will behave cooperatively, and vice versa. In real settings, the provider's willingness will change from one interaction to another due to many reasons, e.g. the change in the cost to provide a high-quality service. Such willingness fluctuation is also simulated in the experiments. Before each experiment, an *initial willingness* is assigned to the provider. The value of the initial willingness is varied to simulate the providers with different nature. In the experiments, initial willingness can be 0.9, 0.5 or 0.1, which corresponds to the provider *highly willing to behave cooperatively*, *willing to behave cooperatively*, and *reluctant to behave cooperatively* respectively². Then in each of the subsequent interactions with the consumers, it can change its willingness. The willingness change each time is always ≤ 0.2 , and the updated value is restricted to remain in the range of $[0, 1]$. Three strategies of willingness change are simulated in the experiments, namely, increasing or decreasing from the one in last interaction, or remaining the same as the one in last interaction. The provider has an equal probability of choosing one of the three fluctuation strategies. That is, the probability that the provider would choose one of the three strategies in each interaction is set to $1/3$. Similar method to simulate providers' behavior has also been applied in research work that studies the performance of Beta Reputation System, e.g [JHF03].

²As the provider's behavior fluctuates around its "initial willingness", we use the value of "initial willingness" to refer to and differentiate various categories of providers.

The purpose of the experiments is to study the proposed filtering method's effectiveness in mitigating the adverse effect of unfair testimonies. To facilitate the study, two types of unfair testimonies are simulated in the experiment:

- **Ballot-stuffing.** When generating the testimony, an agent, with a probability p_{unfair} , changes its rating in each interaction to be *positive* regardless of its real experience with the provider in this interaction. This will make the derived testimony unfairly positive.
- **Badmouthing.** When generating the testimony, an agent, with a probability p_{unfair} , changes its rating in each interaction to be *negative* regardless of its real experience with the provider in this interaction. This will make the testimony unfairly negative.

The consumer's evaluation of trust on the provider is basically a subjective prediction of the probability that the provider would behave cooperatively in future. On the other hand, the probability that provider would behave cooperatively is controlled by its willingness. Therefore, the consumers' evaluations of trust on the provider should reflect its true willingness if the evaluations are not influenced by the unfair testimonies. Hence, the effectiveness of the proposed filtering method can be measured on the basis of the difference between the evaluation of trust and the provider's willingness, which can be quantified as the *mean absolute error* (MAE) :

$$MAE = \frac{\sum_{I \in \mathcal{I}} |TP_{s,o} - L_o|}{\|\mathcal{I}\|} \quad (3.15)$$

where $\mathcal{I}$ is the set of all the interactions, and $TP_{s,o}$ and L_o stand for the consumer agent's evaluation of trust on the provider and the provider's willingness in the corresponding interaction respectively. The average of all consumers' MAE is obtained as final measurement of the effectiveness of the proposed filtering method. A lower value of MAE implies that the proposed filtering method is more effective in mitigating the adverse effect of the unfair testimonies, and vice versa.

In order to study the proposed filtering method's effectiveness in a more objective manner, it is also compared with other related work. The closest related work is Whitby et al.'s proposal in [WJI04]. This work is also an extension of the Beta Reputation System, and also

applies a filtering method to mitigate the adverse effect of the unfair testimonies. As discussed in Section 2.3.2.1, Whitby et al.'s proposal first generates a majority opinion by aggregating all the available testimonies. Then it identifies the possible unfair testimony by testing whether a testimony is outside the q quantile and $(1 - q)$ quantile of the majority opinion. If the test outcome is positive, the testimony is considered as a possible unfair testimony and will be discarded. Then the majority opinion is generated again with the remaining testimonies before a new round of testimonies filtering. In the experimental studies, the quantile parameter is set as $q = 0.01$, since $q = 0.01$ is a good choice as reported in [WJI04].

Comparisons have been conducted to examine these two methods' effectiveness in various configurations. An experiment configuration is made up of three factors, namely the provider's initial willingness (IL), p_{unfair} , and the ratio of unfair testimonies, i.e. percentage of consumers that give unfair testimonies. As discussed earlier in this section, the initial willingness controls the behavior of the provider. p_{unfair} controls the "strength" of each unfair testimony. That is, the higher the p_{unfair} is, the more individual ratings are modified. Consequently, this implies that the unfair testimony deviates more from the trustee agent's real willingness, and vice versa. p_{unfair} takes a value of high (0.9), medium (0.5), or low (0.1). The ratio of unfair testimonies is taken from the set of $\{0.1, 0.4, 0.7, 1\}$. A higher ratio indicates that there will be more agents giving unfair testimonies, and vice versa. Five individual experiments are run with each experiment configuration, and the average of the results obtained in the five experiments is taken as the final result of the corresponding configuration.

3.4.2 The Results

Before presenting the main experimental results, we first show the adverse effect of unfair testimony. Figure 3.3 shows the provider's willingness and one consumer's evaluations of trust on the provider over the 1000 interactions both with and without the presence of unfair testimonies, more specifically, *bad-mouthing* unfair testimonies. In both cases, the initial willingness of the provider is set as 0.9. Figure 3.3 (a) shows the results in configuration with $p_{unfair} = 0.5$ and unfair ratio as 70%.

Figure 3.3: Adverse effect of the unfair testimonies: (a) without unfair testimonies; (b) with badmouthing testimonies

It is observed in Figure 3.3 (a) that the consumer's evaluations of trust are close to the provider's willingness without the presence of unfair testimonies. However, in Figure 3.3 (b), there is an obvious difference between the evaluations of trust and the willingness in the presence of badmouthing unfair testimonies. This observation clearly shows that the presence of unfair testimonies does affect the evaluations.

Figure 3.4: The trust model in the presence of unfair testimonies

Figure 3.4 illustrates the consumer's evaluations of trust on the provider and the provider's willingness in the presence of badmouthing unfair testimonies but with the proposed filtering method applied. Compared with Figure 3.3 (b), a smaller distance between the evaluations of trust and willingness can be observed, which shows that the proposed filtering method does mitigate the adverse effect of the unfair testimonies.

Experiments were then conducted to study the effectiveness of the proposed filtering method in various experimental configurations. The proposed filtering method's effectiveness (in terms of MAE) in different experimental configurations are plotted as lines with labels suffixed by “(O)” in Figures 3.5-3.6. We also measure the effectiveness of our proposal against Whitby et al.'s proposal in [WJI04]. Effectiveness (in terms of MAE) of Whitby et al.'s proposal are plotted in Figure 3.5-3.6 with labels suffixed by “(W)”.

It is observed in Figure 3.5-3.6 that MAE increases (i.e. effectiveness decreases) with the increase of unfair testimonies ratio. Generally, the proposed filtering method achieves a MAE of less than 0.1 when fewer than 50% consumers give unfair testimonies³ (i.e. unfair testimonies ratio $\leq 50\%$). However, when more than 50% consumers give unfair testimonies, an obvious increase of the MAE (i.e. a reduced effectiveness) is observed. This is because the proposed filtering method basically depends on what the public says about the trustee (i.e. the current belief generated by aggregating third-party testimonies) to filter the testimonies. However, even the public opinion cannot reflect the provider's true behavior when more than 50% consumers give unfair testimonies.

It is also noted that MAE increases correspondingly with an increase of p_{unfair} . With the increase of p_{unfair} , each individual rating is more susceptible to be manipulated when the consumer generates the testimonies, which results in the testimonies being more deviant from the provider's real willingness. Thus, the unfair testimonies becomes more powerful in biasing the public opinion. Consequently, the proposed filtering method becomes less effective since it depends on public opinion to filter the testimonies.

3.4.2.1 Comparison with Whitby et al.'s Proposal

Similar increasing trends with the increase of unfair testimony ratio and p_{unfair} are also observed with Whitby et al.'s proposal [WJI04]. Nevertheless, the proposed method still outperforms Whitby et al.'s proposal in most of the configurations that are studied by the experiments.

³There are exceptions, which will be discussed later.

(a) IL=0.1

(b) IL=0.5

(c) IL=0.9

Figure 3.5: Effectiveness in the presence of ballot-stuffing testimonies

(a) IL=0.1

(b) IL=0.5

(c) IL=0.9

Figure 3.6: Effectiveness in the presence of badmouthing testimonies

In configurations where fewer than 50% consumers give unfair testimonies, the effectiveness improvement over Whitby et al.'s proposal is less obvious. When more than 50% consumers give unfair testimonies, the improvement becomes more obvious.

The proposed filtering method is also more effective than Whitby et al.'s proposal in terms of processing time. The proposed method takes about 0.02 second on average to process all the testimonies from the other 99 consumers in each interaction. However, with identical experiments setting (Intel Xeon CPU 3.0GHz and 2GB memory), Whitby et al.'s proposal takes about 1.5 seconds on average. This is primarily due to the iterative nature of Whitby et al.'s proposal, which processes each testimony more than once before the truster consumer derives its final evaluation of trust on the provider. For this reason, it does not scale well with the increase in the number of available testimonies. In contrast, the proposed method processes every testimony only once, which makes it scale linearly with the increase in the number of available testimonies.

Such a disparity in performance is mainly due to the difference in the strategies employed to exploit public opinion. Whitby et al.'s proposal generates public opinion by aggregating testimonies before filtering. The public opinion is already biased by the unfair testimonies, which makes it easier for the unfair testimonies to escape the filtering. In contrast, the proposed filtering method examines each testimony before it is aggregated to generate a new majority opinion. A possible unfair testimony is discarded before it can bias the majority opinion. Only a testimony that is close to (within an acceptable range) the truster's current belief (i.e. the public opinion) is aggregated.

3.5 Discussion

In this chapter, we introduce trust into MAS by exploiting an existing basic trust model, i.e. Beta Reputation System (BRS). As a starting point in this research to tackle the presence of unfair testimonies, this chapter proposes a testimony-filtering method to mitigate the adverse effect of unfair testimonies, and examines its effectiveness in the context of the BRS.

An obvious advantage of the proposed testimony-filtering method is that it works by analyzing the testimonies directly. It does not require any additional mechanism or information used in other

existing methods, such as the incentive method used in Jurca and Faltings' model [JF03, JF04] and the method used in ReGrE_T model [SM03]. We also study the effectiveness of the proposed uncertainty-based filtering method by comparing it with Whitby et al.'s proposal [WJI04], which is a typical *endogenous* filtering method to tackle unfair testimonies. Experimental results show that the proposed filtering method outperforms Whitby et al.'s proposal in terms of both the processing time and the effectiveness to mitigate the adverse effect of the unfair testimonies. Although the effectiveness of the proposed filtering method is studied in the context of the BRS, it can easily be adapted in trust models that use non-binary ratings. The testimony quality can be measured with the same criterion (i.e. Eq. (3.10)) because it employs entropy as the basis of the testimony quality measurement.

Despite the advantages the proposed uncertainty-based filtering method has, there is still room for improvement. For example, the effectiveness of the filtering method depends on the selection of the filtering threshold ε . A careful selection of ε is necessary in order to apply the method to different scenarios. An adaptive criterion is preferred. The proposed filtering method can easily be adapted to trust models that use non-binary ratings. However, this assumes that the witnesses share the raw counts of satisfactory and unsatisfactory interactions directly with the trustees as testimonies. This hampers its wider applicability to other trust models that do not share raw experience as testimonies.

Moreover, we only discuss the observations about the p_{unfair} and unfair testimonies ratio in Section 3.4.2, and leave out the influence of the *initial willingness* (IL) and the type of unfair testimonies. In fact, it can be seen that there is a change in the proposed filtering method's effectiveness when the provider's willingness varies. The observations from Figure 3.5-3.6 show that:

- If the provider's willingness is low, i.e. its willingness is fluctuating around $IL = 0.1$, the proposed filtering method is more effective in the presence of more than 50% *badmouthing* unfair testimonies than in the presence of same ratios of *ballot-stuffing* unfair testimonies when the p_{unfair} is high, e.g. $p_{unfair} \geq 0.5$ (see Figure 3.5(a)).

- If the provider's willingness is high, i.e. its willingness is fluctuating around $IL = 0.9$, the proposed filtering method is more effective in the presence of more than 50% *ballot-stuffing* unfair testimonies than in the presence of same ratios of *badmouthing* unfair testimonies when the p_{unfair} is high, e.g. $p_{unfair} \geq 0.5$ (see Figure 3.6(c)).

These two observations are primarily due to a high value of p_{unfair} , which would “swap” the value of $\Pr(\alpha)$ and $\Pr(\beta)$ in Eq. (3.12). This makes the unfair testimonies' qualities very close to, or even the same as, the consumer's current belief about the provider. For example, suppose the consumer's current belief closely reflects the provider's willingness and indicates that the provider has provided satisfactory services in 8 out of 10 past interactions. In this case, $\Pr(\alpha) = (8 + 1)/(10 + 2) = 0.75$, and $\Pr(\beta) = (2 + 1)/(10 + 2) = 0.25$. A consumer with high value of p_{unfair} giving badmouthing testimonies may report that the provider has provided unsatisfactory services in 8 out of 10 past interactions, which makes $\Pr(\alpha) = 0.25$ and $\Pr(\beta) = 0.75$. Although the values of $\Pr(\alpha)$ and $\Pr(\beta)$ are swapped in these two cases, both the qualities of the consumer's current belief and the badmouthing testimony are $-0.75 \log(0.75) - 0.25 \log(0.25)$. Thus, the *badmouthing* unfair testimony escapes the test in Eq. (3.14). For the same reason, the *ballot-stuffing* unfair testimonies with high p_{unfair} would escape the filtering when the provider's willingness is low, i.e. willingness fluctuating around $IL = 0.1$.

As these observations show, to mitigate the adverse effect of unfair testimonies, only measuring the quality of an individual testimony's based on the degree of uncertainty does not always suffice. A possible alternative is to measure the quality (or usefulness) of the witness' testimony based on how much this witness' testimonies has reduced the truster's uncertainty regarding the provider's behavior. The rationale here is that if the witness' testimonies are useful in reducing the uncertainty of the truster in the past, the same witness' new testimonies should also be useful too. Such a measurement of the quality (or usefulness) of third-party testimony is consistent with the effect of third-party testimonies. For instance, as pointed out in [BK95], the effect of third-party testimonies is to make the truster more certain of its trust on the interaction partners.

Inspired by the alternatives to improve the proposed uncertainty-based testimony-filtering method, the following chapter revamps it with a credibility model.

Chapter 4

Coping with Unfair Testimonies: A General Credibility Model

Chapter 3 proposes an uncertainty-based testimony-filtering method to mitigate the adverse effect of unfair testimonies. Despite the merits that the proposed uncertainty-based filtering method has, there is still room for improvement. This chapter revamps this filtering method with a novel credibility model. The proposed credibility model measures the witness agent's credibility based on the usefulness of its past testimonies, and then filters and aggregates the testimonies according to corresponding witness' credibility.

4.1 Introduction

In Chapter 3, an uncertainty-based testimony-filtering method is proposed to address the adverse effect of unfair testimonies. A potential improvement is to measure the quality or usefulness of a witness's testimonies based on how much the witness's testimonies have reduced the truster's uncertainty regarding the trustee's behavior.

Measuring such reduction of uncertainty requires the witness's past testimonies on the trustees and the truster's own evaluations of trust on the corresponding trustees. However, in the basic trust model and the filtering method presented in Chapter 3, the testimonies reported by a witness as well as the truster's own evaluation of trust on the trustees are not stored after they are aggregated to evaluate the overall trust on the trustees. This makes it impossible to calculate such reduction of uncertainty.

To solve this problem, we revamp the uncertainty-based filtering method with a *credibility* model. The credibility model has a profile-building component, which records the information about witness agent's past testimonies, as well as the truster's own past evaluations of trust on the corresponding trustees. Then a credibility metric is designed to measure the witness's credibility, i.e. how much the witness's testimonies have reduced the truster's uncertainty about the trustees. As discussed in Section 1.1.1, the usefulness of a witness's testimony can be measured based on how much the witness has reduced the truster's uncertainty in the past with its successful testimonies-reporting. Intuitively, if a witness's testimonies are useful in the past in reducing the truster's uncertainty, its new testimony is also likely to be useful. Such a measurement of usefulness is defined as the witness's *credibility* in this thesis. Finally, the testimonies are filtered according to the credibility of corresponding witnesses. Furthermore, the remaining testimonies are aggregated with different weights according to the corresponding witnesses' credibility to further reduce the adverse effect of unfair testimonies.

It is noted in this thesis that *trust* and *credibility*, though related, are considered as two different concepts. The truster's evaluation of trust on the trustee captures its uncertainty about the trustee's behavior. However, the truster's evaluation of an agent's (i.e. witness's) credibility is a subjective evaluation of the usefulness of the latter's testimonies in reducing the uncertainty. Correspondingly, the *trust model* refers to the model that is used by the agent to evaluate trust on the trustee, whereas the *credibility model* refers to the model that is used to evaluate the credibility of witness agent. The proposed credibility model does not explicitly require a specific trust model to be employed by the agents. The proposed credibility model is therefore applicable to a wide range of trust models, though this thesis continues to use the basic trust model presented in Section 3.2.

This chapter presents the design of the proposed credibility model. Section 4.2 introduces the notations used in this chapter. Section 4.3 presents the profile-building component of the credibility model. After that, the design of the credibility metric is presented in Section 4.4. Section 4.5 discusses the component to filter and aggregate the testimonies. Finally, Section 4.7 summarizes this chapter.

4.2 Notations

Other than the notations defined in the previous chapter (see Section 3.2.1), we introduce some new notations for the ease of presentation.

Before each interaction with trustee agent a_o , truster a_s will evaluate its trust on the trustee. The trust evaluation is a combination of a_s 's own *direct trust* on a_o and third-party testimonies. The direct trust on a_o is derived based on truster a_s 's personal interaction experiences with a_o , and is denoted as $TP_{s,o}$. The trust evaluation with a combination of $TP_{s,o}$ and third-party testimonies is called truster a_s 's *overall trust* in a_o , which is denoted as $TA_{s,o}$.

$TA_{s,o}$ is the *pre-interaction* trust evaluation in that it is calculated before truster agent a_s interacts with trustee agent a_o . If truster agent a_s determines to interact with trustee a_o after the pre-interaction trust evaluation, it will evaluate the trustee's trustworthiness again with new observation of the trustee's behavior in the interaction. This is called the *post-interaction* trust evaluation, and it contains only the truster's direct trust on the trustee, i.e. it is made only with truster's personal interaction experience with the trustee (including the new experience in the interaction).

After the truster obtains the post-interaction trust evaluation, it will update the profiles of those witnesses who have reported testimonies when making the pre-interaction evaluation, and re-evaluates the witnesses' credibility accordingly. The truster agent a_s 's evaluation of a witness agent a_w 's credibility is denoted as $TT_{s,w}$.

We reiterate that, *trust* and *credibility* are two different concepts, though related. Truster a_s 's evaluation of *trust* in trustee a_o , i.e. $TP_{s,o}$ or $TA_{s,o}$ captures a_s 's uncertainty about a_o 's future behavior. However, a_s 's evaluation of agent a_w 's (as a witness of agent a_o) *credibility*, i.e. $TT_{s,w}$ is a subjective evaluation of the usefulness of a_w 's testimonies in reducing the uncertainty.

4.3 Maintaining the Witness' Profile

The credibility model contains three primary components: (1) a component used by the truster to maintain a profile for the witness agent; (2) a metric with which the truster evaluates the

witness agent's credibility; and (3) a component with which the truster filters and aggregates the testimonies. This section discusses the profile maintenance. The other components are discussed in the subsequent sections.

4.3.1 Preprocessing the Testimonies

With the decoupling of credibility and trust, the proposed credibility model does not require any specific trust model to evaluate the trust. However, the basic trust model assumes that the witnesses share the raw counts of satisfactory and unsatisfactory interactions with the trustees as testimonies. This form of testimony is specific to the *Beta Reputation System*, whereas there exist trust models that do not apply this form of testimonies. Therefore, we discard this assumption by changing the form of testimonies. That is, instead of sharing the counts of positive and negative experience directly as testimony, the witness's evaluation of the direct trust on the trustee is shared as testimony. Such form of testimonies is commonly-used in most of the existing trust models. For example, all the models reviewed in Chapter 2 other than the model proposed by Abdul-Rahman and Hailes (see Section 2.3.3.1) employ continuous ratings to represent the outcome of trust evaluations. It is also consistent with the nature of trust as the subjective probability that the trustee would behave cooperatively.

We have seen the advantages of changing the form of testimonies. However, due to trust's probabilistic nature, the testimony usually takes continuous values in the range of $[0, 1]$ after the changing, thus requiring more overhead than in the basic trust model in maintaining and storing the testimonies. In the basic trust model, the testimonies can easily be maintained and stored by only updating the counts of positive and negative ratings (see Section 3.2). To address this issue, the credibility model pre-processes the continuous testimonies to become discrete ratings with a scale of Z before they are stored in the witness's profile. The preprocessing retains the benefit of maintaining and storing the testimonies in the basic trust model, which will be discussed in the next subsection. The preprocessing is done by (1) choosing a scale for the discrete ratings (denoted as Z), (2) stratifying the continuous range of $[0, 1]$ into Z bins, and (3) representing the testimonies using the indices of the corresponding bins which contain the testimonies.

4.3.2 The Organization of the Profile

After the preprocessing, agent a_s maintains the profile for a witness agent a_w in the form of a contingency table as shown in Figure 4.1. It is basically an alignment of the agent a_w 's past testimonies and the truster a_s 's own post-interaction evaluations of direct trust on the corresponding trustees. It is noted that, the truster will use a table to maintain the witness agent's past testimonies on all trustees.

a_s 's post-interaction direct trust	a_w 's testimonies					Total
	1	...	j	...	Z	
1	n_{11}	...	n_{1j}	...	n_{1Z}	R_1
$\vdots$			$\vdots$			$\vdots$
i	n_{i1}	...	n_{ij}	...	n_{iZ}	R_i
$\vdots$			$\vdots$			$\vdots$
Z	n_{Z1}	...	n_{Zj}	...	n_{ZZ}	R_Z
Total	C_1	...	C_j	...	C_Z	N

Figure 4.1: Agent's profile of giving testimonies

In Figure 4.1, Z denotes the scale of the ratings used to represent the testimonies¹. N is the number of successful testimony-reporting by witness agent a_w . A successful testimony-reporting means that (1) the testimony reported by agent a_w has been used by a_s in the pre-interaction trust evaluation, and (2) a_s interacts with that trustee after the pre-interaction trust evaluation. If condition (1) is met but condition (2) is not, the testimony-reporting is not successful, and it will not be recorded in the profile.

Each cell n_{ij} records the number of successful testimony-reporting in which a_w 's testimonies are j while a_s 's post-interaction direct trust on the same trustees are i . $n_{ij} = 0$ if there is no such testimony-reporting. R_i records the total count of testimony-reporting in which agent a_s 's post-interaction direct trust on the trustees are i , while C_j records the total count of testimony-reporting in which agent a_w 's testimonies are j . In other words, R_i is the sum of the cells in the

¹This scale and the scale of trust information that are used to evaluate trust are not related. For example, a trust model can use binary ratings to evaluate trust, while the outcomes of the trust evaluation (i.e. the testimonies) can be discretized to ratings with a scale of 5.

row i , and C_j is the sum of the cells in the column j , i.e.

$$R_i = \sum_{j=1}^Z n_{ij}, \quad C_j = \sum_{i=1}^Z n_{ij}, \quad \sum_i R_i = \sum_j C_j = N \quad (4.1)$$

Preprocessing the testimonies to discretize the ratings makes the update of the profile easier. That is, when the witness makes a new successful testimony-reporting, the profile can be updated by just increasing one of the cells in the profile instead of storing the testimony directly. For example, witness agent a_t has reported a testimony as 0.25 to truster agent a_s on trustee agent a_o , while a_s 's own post-interaction direct trust on a_o is 0.15. Suppose the testimonies are discretized into $Z = 5$ bins, a_t 's testimony and a_s 's post-interaction direct trust are pre-processed to discrete ratings 2 and 1 respectively. Correspondingly, the profile can be updated by increasing the value of cell n_{12} by 1. Moreover, since a witness's credibility will be measured based on its profile, discrete values in its profile are generally more computationally tractable than continuous values.

An example of the profile is shown below.

Example 4.1 *Suppose the testimonies are discretized to ratings with a scale of 5. A witness agent a_w has made 3 successful testimony-reporting to truster agent a_s :*

- (i) *Agent a_w gave 2 testimonies of 5 on two trustees, and a_s 's post-interaction direct trust on the corresponding trustees are 3, i.e. $n_{35} = 2$.*
- (ii) *Agent a_w gave another testimony of 4 on a trustee, and a_s 's post-interaction direct trust on the corresponding trustee is 4, i.e. $n_{44} = 1$.*

After these 3 successful testimony-reporting, a_s maintains the profile for agent a_w as shown in Figure 4.2. ■

4.4 Measuring the Credibility

On the basis of the information recorded in the profile, credibility of the witness agent can be measured.

a_s 's post-interaction direct trust	a_w 's testimonies					Total
	1	2	3	4	5	
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	2	2
4	0	0	0	1	0	1
5	0	0	0	0	0	0
Total	0	0	0	1	2	3

Figure 4.2: Agent a_w 's profile after 3 successful testimony-reporting

4.4.1 The Derivation of the Credibility Metric

Suppose, a_w 's testimonies are the only information available for truster agent a_s to make the pre-interaction trust evaluations, and agent a_s has built a profile of witness agent a_w to capture a_w 's past testimony-reporting. The truster agent a_s 's pre-interaction evaluations of trust on the trustees can be considered as an activity to predict its post-interaction direct trust on the trustee in the following two cases [GK54]:

Case (1): a_s makes the evaluations given no testimony from witness a_w .

Case (2): a_s makes the evaluations given a_w 's testimonies.

In **Case (1)**, since agent a_s has the information about its past post-interaction direct trust, it cannot be worse off by predicting its post-interaction direct trust on the trustee based on its past experiences in interacting with other trustees, even though a_w has no testimony. More specifically, a_s predicts its direct trust on the trustee as a particular rating i with a probability R_i/N . By doing so, the proportion of the consistent predictions in the long run is $\sum_i (R_i/N)^2$ [GK54]. Here, a prediction is consistent when the truster's pre-interaction evaluation of trust is equal to its post-interaction evaluation of direct trust on the corresponding trustee².

In **Case (2)**, in light of witness a_w 's testimony, the probability that a_s predicts its direct trust on the trustee as a particular rating i is essentially the posterior probability of i 's occurrence given

²Eq. (3.15) measures the quality of trust evaluation by calculating the distance between the trust evaluation and the trustee's real behavior. Similarly, the distance between truster's pre-interaction trust evaluation and the trustee's real behavior is measured here. However, because the credibility model has pre-processed the trust evaluations (i.e. testimonies) into discrete ratings, the distance is measured by determining whether the two ratings are same, which is interpreted as consistency here.

a_w 's testimony of j 's presence in the profile. If we denote the prior probability of testimony j 's occurrence as $P(j)$ and the posterior probability of i 's occurrence given j as $P(i|j)$, according to Bayes' theorem, we have

$$P(i|j) = \frac{P(j|i)P(i)}{P(j)} = \frac{\frac{n_{ij}}{R_i} * \frac{R_i}{N}}{\frac{C_j}{N}} = \frac{n_{ij}}{C_j}$$

In other words, a_s predicts its direct trust on the trustee as i with a probability as $(\frac{n_{ij}}{N})/(\frac{C_j}{N})$. In this case, the proportion of the consistent predictions in the long run is $\sum_i \sum_j (\frac{n_{ij}}{N})^2 / (\frac{C_j}{N}) = \frac{1}{N} \sum_i \sum_j (n_{ij}^2 / C_j)$ [GK54].

Recall that credibility of the witness is measured by how much its past testimonies have reduced the truster's uncertainty regarding the trustees' behavior. Such a reduction can be quantified as the reduction of a_s 's inconsistent predictions given the witness's testimonies in the past, i.e. the reduction of inconsistent predictions as a_s goes from **Case (1)** to **Case (2)**. Therefore, $TT_{s,w}$ is evaluated according to the following formula:

$$\begin{aligned} TT_{s,w} &= \frac{(1 - \sum_i (R_i/N)^2) - (1 - \frac{1}{N} \sum_i \sum_j (n_{ij}^2 / C_j))}{1 - \sum_i (R_i/N)^2} \\ &= \frac{N \sum_i \sum_j \frac{n_{ij}^2}{C_j} - \sum_i R_i^2}{N^2 - \sum_i R_i^2} \end{aligned} \quad (4.2)$$

$TT_{s,w}$ takes values in the range of $[0, 1]$. $TT_{s,w} = 0$ implies that agent a_w 's testimonies has not reduced a_s 's uncertainty at all. $TT_{s,w} = 1$ implies a_s 's perfect predictability of direct trust on the trustees given witness a_w 's testimonies.

Note that different trusters may have different evaluations of credibility of the same witness. This is because the credibility is basically a subjective measurement of the usefulness of the witness's testimonies, and different trusters may have different experience in receiving testimonies from the same witness.

4.4.2 Bootstrapping of the Profile and Credibility

As mentioned in Section 4.3, each cell n_{ij} in the profile should intuitively be initialized to 0. However, it is easily derived from Eq. (4.2) that this leads to $TT_{s,w} = 0/0$ after witness a_w 's first successful testimony-reporting to agent a_s .

a_s 's post-interaction direct trust	a_w 's testimonies					Total
	1	...	j	...	Z	
1	0	...	0	...	0	0
$\vdots$			$\vdots$			$\vdots$
i	n_{i1}	...	n_{ij}	...	n_{iZ}	R_i
$\vdots$			$\vdots$			$\vdots$
Z	0	...	0	...	0	0
Total	n_{i1}	...	n_{ij}	...	n_{iZ}	R_i

Figure 4.3: A special case leading to 0/0 when evaluating credibility

In order to avoid the undefined 0/0, *dummy cell* is introduced to initialize the profile. That is, every cell in the profile is initialized with a dummy value of $1/Z$ instead of 0. Then all the dummy cells in the j th column are cleared by replacing the dummy cells with 0 after agent a_w 's first successful testimony-reporting with a testimony as j . After the dummy cells are cleared, the corresponding cell in the j th column is updated with real count of successful testimony-reporting accordingly.

There is another case leading to the undefined 0/0 even with the introduction of dummy cell. That is, as Figure 4.3 shows, each column in the profile has only one non-zero cell, and all the non-zero cells are in the same row. This represents the fact that a_s always has only one possible prediction in light of a_w 's testimonies. This implies a_s 's perfect predictability of its post-interaction direct trust given a_w 's testimonies. Consequently, $TT_{s,w} = 1$ in this case, as discussed before.

After initializing the profile with the dummy cells, according to Eq. (4.2), the initial value of $TT_{s,w}$ is 0. However, $TT_{s,w} = 0$ means that witness a_w 's testimonies are not useful at all. It implies that testimony reported by a_w would not be aggregated. This consequently makes $TT_{s,w}$ remain as 0 as the credibility is measured based on a_w 's past testimony-reporting. Therefore, a non-zero initial credibility is necessary to *bootstrap* the testimony aggregation and future computation of the credibility. Moreover, the gap between the initial credibility and its future values should be as small as possible to smooth the credibility evolution.

To this end, $TT_{s,w}$ is initialized to be close to the updated value of $TT_{s,w}$ after a_w 's first successful testimony-reporting to agent a_s . After a_w 's first successful testimony-reporting, there

a_s 's post-interaction direct trust	a_w 's testimonies							Total
	1	...	$q-1$	q	$q+1$	...	Z	
1	$1/Z$	...	$1/Z$	0	$1/Z$	...	$1/Z$	$\frac{Z-1}{Z}$
$\vdots$				$\vdots$				$\vdots$
$p-1$	$1/Z$	...	$1/Z$	0	$1/Z$	...	$1/Z$	$\frac{Z-1}{Z}$
p	$1/Z$	...	$1/Z$	1	$1/Z$	...	$1/Z$	$\frac{Z-1}{Z} + 1$
$p+1$	$1/Z$	...	$1/Z$	0	$1/Z$	...	$1/Z$	$\frac{Z-1}{Z}$
$\vdots$				$\vdots$				$\vdots$
Z	$1/Z$	...	$1/Z$	0	$1/Z$	...	$1/Z$	$\frac{Z-1}{Z}$
Total	1	...	1	1	1	...	1	Z

Figure 4.4: Agent a_w 's profile after its first successful testimony-reporting

is only one non-zero cell in a_w 's profile. Assume that the non-zero cell is in the p th row and q th column, i.e. $n_{pq} = 1$. a_w 's profile is updated by a_s as shown in Figure 4.4. By applying Eq. (4.2) on the updated profile, it is obtained that $TT_{s,w} = 1/(Z+1)$. Therefore, agent a_s initializes agent a_w 's credibility as $TT_{s,w} = 1/(Z+2)$, since $1/(Z+2)$ is close to $1/(Z+1)$. Moreover, $1/(Z+2) < 1/(Z+1)$. This corresponds to the intuition that a_w is more credible after its first testimony-reporting to a_s since its testimony has reduced a_s 's uncertainty.

4.4.3 Evolution of the Credibility

Since a_w 's credibility is measured based on its past testimonies, $TT_{s,w}$ is updated when a_w makes more successful testimony-reporting to agent a_s . Update of $TT_{s,w}$ is triggered each time after agent a_s obtains its post-interaction direct trust on the trustee, i.e. its uncertainty about the trustee has been updated in light of the trustee's new behavior in the interaction. Suppose witness agent a_w 's testimony on trustee agent a_o for a_s 's pre-interaction trust evaluation is j , and a_s 's post-interaction direct trust on a_o is i . The update of $TT_{s,w}$ is carried out with the procedure shown in Figure 4.5.

4.5 Testimony Filtering and Aggregation

Before aggregating the testimonies to obtain the overall trust that it should place in trustee a_o , agent a_s will evaluate the credibility of all the witnesses that have reported testimonies on a_o . Then it will choose a subset of testimonies to be aggregated. Unlike the filtering method presented

Procedure $TT_Update(s, w)$ a_s : the truster agent, a_o : the trustee agent, a_w : a witness of the trustee agent a_o .

```

1: if  $a_w$  is a newly-encountered witness for  $a_s$  then
2:    $a_s$  initializes the profile to record  $a_w$ 's testimony-reporting
3:   for  $c = 1$  to  $Z$  do
4:     for  $r = 1$  to  $Z$  do
5:        $n_{rc} = 1/Z$  {fill in the dummy cell}
6:     end for
7:   end for
8: end if
9: if  $a_s$  interacts with  $a_o$  after the pre-interaction trust evaluation then
10:   $a_s$  evaluates the post-interaction direct trust on  $a_o$  with new observation of  $a_o$ 's behavior
11:  if  $a_w$  makes a successful testimony-reporting with a testimony of  $j$  for the first time then
12:    for  $r = 1$  to  $Z$  do
13:       $n_{rj} = 0$  {clear the dummy cell in  $j$ th column}
14:    end for
15:  end if
16:   $n_{ij} = n_{ij} + 1$ 
17:   $a_s$  re-evaluates  $TT_{s,w}$  with Eq. (4.2) based on the updated profile
18: end if

```

Figure 4.5: Procedure to update credibility

in Chapter 3 in which the testimonies are filtered based on the “uncertainty” that the testimonies contain, a_s now filters the testimonies according to the corresponding witnesses’ credibility with the credibility model deployed.

The filtering of the testimonies is done by comparing the corresponding witness’s credibility against the *confidence* of truster agent a_s ’s own evaluation of direct trust on the trustee. In order to evaluate the confidence, truster agent a_s also maintains a profile for itself, which is organized in the same form as Figure 4.1. Nevertheless, the cells are attached with different meaning in this case. Each cell n_{ij} in the profile that agent a_s builds for itself records the number of interactions with the trustees in which a_s ’s pre-interaction direct trust are j while its post-interaction direct trust are j . Then the confidence is derived with Eq. (4.2) according to the profile, and it is denoted as $TT_{s,s}$.

Then the testimony reported by a witness agent a_w will be aggregated if:

$$TT_{s,w} \geq TT_{s,s} \quad (4.3)$$

where $TT_{s,w}$ is agent a_w 's credibility, and $TT_{s,s}$ is agent a_s 's confidence about its own direct trust evaluations. Unlike the filtering method presented in Chapter 3, which filters the testimonies based on a predefined filtering threshold ε , this filtering criterion is more adaptive, in that only those testimonies that are more useful than a_s 's own evaluations of direct trust are aggregated.

Then the selected testimonies will be aggregated by truster a_s . A conservative view towards the testimonies is employed in the testimony aggregation. That is, each testimony is considered as possible unfair testimony in first place, and it is adjusted afterward to reduce the adverse effect of the possible unfair testimonies among the selected testimonies.

The testimony is adjusted based on the corresponding witness's previous testimony-reporting recorded in its profile. Suppose agent a_w gave a testimony of j to truster agent a_s . According to **Case (2)** in Section 4.4.1, given witness a_w 's testimony j , agent a_s predicts its direct trust on the trustee as v with a probability $\frac{n_{vj}}{N} / \frac{C_j}{N} = n_{vj} / C_j$. Or put it in another way, witness a_w 's testimony j is adjusted to rating v with a probability n_{vj} / C_j , i.e.:

$$\Pr(j \rightarrow v) = n_{vj} / C_j \quad (4.4)$$

The rationale of the testimony adjustment is that: if, in the past, each time witness a_w gives a testimony as j , truster a_s 's post-interaction direct trust on the trustee is often found to be v . Then, most probably, its direct trust on the trustee would be v if a_w gives a testimony as j again. The testimony adjustment is important here because it keeps the testimony aggregation consistent with the derivation of the credibility metric (see Case (2) in Section 4.4.1). Moreover, it helps to reduce the adverse effect of the possible unfair testimonies even they are aggregated.

It is possible that truster a_s receives a testimony from a newly-encountered witness. In this case, this testimony is used directly without adjustment as there is no information recorded in the profile to support the adjustment.

An example of the testimony adjustment is given below.

Example 4.2 *After the first 3 successful testimony-reporting in Example 4.1, agent a_w makes another 2 successful testimony-reporting to truster agent a_s : a_w gives two testimonies as 4 and 5 respectively on two trustees, and a_s 's post-interaction direct trust on the two corresponding*

trustees are 3. Witness agent a_w 's profile after 5 successful testimony-reporting is shown in Figure 4.6:

a_s 's post-interaction direct trust	a_w 's testimonies					Total
	1	2	3	4	5	
1	1/5	1/5	1/5	0	0	3/5
2	1/5	1/5	1/5	0	0	3/5
3	1/5	1/5	1/5	1	3	23/5
4	1/5	1/5	1/5	1	0	8/5
5	1/5	1/5	1/5	0	0	3/5
Total	1	1	1	2	3	8

Figure 4.6: Agent a_w 's profile after 5 successful testimony-reporting

Assume a_w gives a new testimony to a_s , and it is 4. Then, before a_s aggregates this new testimony, it does the following testimony adjustment:

$$\Pr(4 \rightarrow 3) = n_{34}/C_4 = 1/2, \quad \Pr(4 \rightarrow 4) = n_{44}/C_4 = 1/2 \quad (4.5)$$

That is, it adjusts this testimony to 3 with a probability as $\frac{1}{2}$, and adjusts to 4 with a probability as $\frac{1}{2}$.

If a_w 's new testimony is 5, then it is adjusted to 3 with a probability as $\Pr(5 \rightarrow 3) = n_{53}/C_5 = 3/3 = 1$. ■

Then the testimonies (after filtering and adjustment), including a_s 's own pre-interaction evaluation of direct trust on the trustee, are aggregated to obtain a_s 's overall trust. The aggregation is done as a weighted mean of the testimonies (after adjustment) with each individual witness's credibility as the weight:

$$TA_{s,o} = \frac{\sum_{a_w \in \mathcal{W}_{s,o}} (TT_{s,w} * TP'_{w,o})}{\sum_{a_w \in \mathcal{W}_{s,o}} TT_{s,w}} \quad (4.6)$$

Here $\mathcal{W}_{s,o}$ is a set of witness agents whose testimonies will be aggregated, including agent a_s itself. $TP'_{w,o}$ is testimony (after adjustment) by agent a_w , who is one of the witnesses in $\mathcal{W}_{s,o}$. Each testimony is given a weight as the corresponding witness's credibility. The testimonies from less credible witnesses will have a smaller impact on the aggregation, and vice versa. Truster a_s 's own pre-interaction direct trust on a_o is aggregated without adjustment, and is assigned a weight as its confidence $TT_{s,s}$.

$TA_{s,o}$ obtained with Eq. (4.6) is in the range of $[1, Z]$, since the testimonies have been pre-processed to become discrete ratings with a scale of Z . In scenarios where the truster has a list of candidate partners from which it chooses one to interact with, this result can be used directly. This is because usually a rational agent would choose to interact with the one with the highest trustworthiness, and the results derived with Eq. (4.6) are enough to differentiate the candidate partners. However, in scenarios where the truster needs to determine whether to interact with a specific partner, $TA_{s,o}$ obtained with Eq. (4.6) may need to be converted to the range of $[0, 1]$, since the decision may be made based on a threshold on the probability that the trustee agent would behave cooperatively.

Figure 4.7 summarizes the approach taken by the proposed credibility model.

4.6 Experimental Study

Experiments have been conducted to study the proposed credibility model's performance in mitigating the adverse effect of unfair testimonies.

First of all, a set of experiments are conducted to compare the performance of the credibility model against the uncertainty-based filtering method discussed in Chapter 3.

Experiments with a publicly available dataset have also been conducted to study the credibility model's effectiveness. The dataset used is the MovieLens dataset³, which contains agents' opinions on a number of movies. By using this dataset, the movies become the trustees, while the trusters are the agents. As discussed before in Section 4.3.1, the credibility model does not concern about which specific model is used to evaluate the trustees since it is decoupled from the trust model. Therefore, only the trusters' evaluations of the trustees matter, which are captured in the dataset. The dataset is used to simulate the exchange of testimonies among the agents. Testimonies are exchanged among agents before they interact with the potential trustees (i.e. whether to watch the movies). Some agent may not tell the truth when giving testimonies. And the credibility model is used to address the presence of unfair testimonies.

³MovieLens dataset is publicly available at <http://www.cs.umn.edu/research/GroupLens/data/>.

The credibility model a_s : the truster agent, a_o : the trustee agent, a_w : a witness of the trustee agent a_o , $TT_{s,w}$: agent a_s 's evaluation of witness a_w 's credibility, $TP_{s,o}$: agent a_s 's pre-interaction direct trust on agent a_o , $TA_{s,o}$: agent a_s 's (pre-interaction) evaluation of overall trust that it should place in trustee a_o , $TP'_{s,o}$: agent a_s 's post-interaction direct trust on agent a_o with new observation of a_o 's behavior in the interaction.

-
- 1: a_s evaluates $TP_{s,o}$ {agents can employ any computational trust model in making the trust evaluation}
 - 2: a_s asks for testimonies from other agents in MAS
 - 3: **for all** witness agent a_w **do**
 - 4: evaluate a_w 's credibility $TT_{s,w}$ using Eq. (4.2)
 - 5: **if** $TT_{s,w} \geq TT_{s,s}$ **then**
 - 6: a_s adjusts a_w 's testimony according to Eq. (4.4) { a_w 's testimony will be aggregated, and is adjusted before being aggregated}
 - 7: **else**
 - 8: a_w 's testimony will not be aggregated
 - 9: **end if**
 - 10: **end for**
 - 11: let $\mathcal{W}_{s,o}$ = all the witnesses whose testimonies will be aggregated
 - 12: a_s aggregates all the adjusted testimonies and $TP_{s,o}$ using Eq. (4.6) to evaluate $TA_{s,o}$
 - 13: **if** agent a_s interacts with the trustee agent a_o after the pre-interaction trust evaluation **then**
 - 14: a_s evaluates $TP'_{s,o}$
 - 15: **for all** witness agent $a_w \in \mathcal{W}_{s,o}$ **do**
 - 16: a_s updates a_w 's credibility $TT_{s,w}$ with $TT_Update(s, w)$ as shown in Figure 4.5
 - 17: **end for**
 - 18: **end if**
-

Figure 4.7: The credibility model

4.6.1 Methodology

For the first set of experiments, same methodology as in Section 3.4 is used. *Mean absolute error* (MAE) (see Eq. (3.15)) is applied again to evaluate the performance.

For the second set of experiments, the MovieLens dataset has already pre-processed the trustees' evaluations of the trusters into discrete ratings of 5-point scale. The dataset contains 100,000 entries of ratings about 1682 items by 943 agents. Each entry has four main fields: the ID of the agent who gives the rating, the ID of the item, the rating, and the timestamp when the rating was given.

The whole dataset is split into two subsets with 80,000 and 20,000 ratings respectively. Ratings in the first subset are taken as truster agents' previous experience with the trustees, based on which each truster initializes the profile for every witness. The latter subset is used to simulate the interaction among agents in the experiments. The experiments are conducted in the following manner:

In each single cycle of the experiment, one rating is taken out by the ascending order of the rating timestamp. In each cycle, the real rating is masked intentionally. The corresponding truster (as indicated by the agent ID field in the entry) needs to ask for testimonies from other witnesses and evaluate the trustee (i.e. evaluate the rating of the movie). The truster usually only interacts with the trustee (i.e. watch the movie) when its evaluation of the trustee is higher than a pre-defined threshold. In order to examine the credibility model's effectiveness extensively, its performance in every single cycle is taken into account by assuming that the threshold is always 0. After that, the truster's evaluation of the trustee is then compared against the trustee's real rating (i.e. the rating that is previously masked) to measure the effectiveness of the proposed credibility model.

Two groups of experiments are conducted. The first group is to study the evaluation accuracy without the presence of unfair testimonies. Similar as what we have done in Section 3.4, *mean absolute error* (MAE) is applied to evaluate the accuracy. Since each truster makes the evaluations individually, it is more reasonable to measure each truster's trust evaluation separately. For this reason, we employ a different interpretation of the MAE metric in this section, which first calculates the mean evaluation error for each truster and then averages all trusters' mean errors :

$$MAE = \frac{\sum_{a_s \in \mathcal{A}_s} \frac{\sum_{a_o \in \mathcal{A}_{so}} (|TP_{s,o} - L_o|)}{\|\mathcal{A}_{so}\|}}{\|\mathcal{A}_s\|} \quad (4.7)$$

where $\mathcal{A}_s$ is the set of all trusters, a_s is one truster in the set of $\mathcal{A}_s$, $\mathcal{A}_{so}$ is the set of trustees that a_s has evaluated, a_o is a trustee in the set of $\mathcal{A}_{so}$, $TP_{s,o}$ and L_o are the a_s 's evaluation of trustee a_o and the trustee a_o 's real rating respectively. A smaller value of MAE is expected to show the credibility model's effectiveness.

The second group is to study the proposed credibility model's effectiveness in the presence of unfair testimonies. Another metric, i.e. *power of unfairness* (POU), is employed as the evaluation metric, which is given by:

$$POU = \frac{\sum_{a_s \in \mathcal{A}_s} \frac{\sum_{a_o \in T_{s,o}} (|r_{\text{target}} - TP_{s,o}| - |r_{\text{target}} - TP'_{s,o}|)}{\|T_{s,o}\|}}{\|\mathcal{A}_s\|} \quad (4.8)$$

where $\mathcal{A}_s$ is the set of all trusters, a_s is one truster in the set of $\mathcal{A}_s$, $T_{s,o}$ is the set of trustees that whose ratings have been manipulated by the unfair testimonies, a_o is a trustee in the set of $T_{s,o}$, and $TP_{s,o}$ and $TP'_{s,o}$ are truster a_s 's evaluation of the trustee a_o with and without the presence of unfair testimonies respectively. r_{target} is the target rating of the unfair testimonies. In the experiments, $r_{\text{target}} = 1$ if the unfair testimonies are to deteriorate the target trustees' trustworthiness (i.e. to lower the target movies' ratings to the lowest rating), or $r_{\text{target}} = 5$ if the unfair testimonies are to push the target trustees' trustworthiness (i.e. to push the target movies' ratings to the highest rating). POU metric measures the extent that proposed credibility model has reduced the effect of the unfair testimonies. A positive value of POU means that the trusters' evaluations have been manipulated towards the target rating by the presence of unfair testimonies. Otherwise, POU gives a negative value. $POU = 0$ if there is no difference in evaluations with and without the presence of unfair testimonies.

The ratio of agents that are giving unfair testimonies is varied to study the proposed credibility model's effectiveness in different settings. The ratios of agents giving unfair testimonies are tuned to be 25%, 50%, and 75%. Experiments with the same settings have been run three times, and the average of the results derived in all the three experiments is taken as the final result.

This section also conducts the comparison study against other credibility modeling methods, in particular the method proposed by Yu and Singh⁴. Instead of initializing the witness' credibility as 1⁴, each truster agent initializes the witness' credibility as the similarity between the two agents' past ratings, which is calculated as the Pearson correlation efficient [LR04] between the two agents' past evaluations on the trustees in the first sub-dataset (with 80,000 rating entries)⁵.

⁴See Appendix A for details about this method.

⁵It is noted that Pearson correlation efficient between two agents could be negative. Those witnesses with negative Pearson correlation efficient are ignored by the trusters when asking for testimonies in the experiments.

Then each witness' credibility is updated using a variant of the Weighted Majority Algorithm⁴. Since all the trustees a truster encounters are new to the truster (the 2nd sub-dataset basically contains the movies that the truster has not rated before), the truster's confidence is always 0⁴. Therefore, the evaluations of the trustees are made by only aggregating all the available testimonies. For both methods, it is implemented that the trusters ask for testimonies from all the witnesses.

In all the experiments presented in this section, the proposed credibility model is implemented with $Z = 10$.

4.6.2 Results

First of all, the effectiveness of the credibility model is compared against that of the uncertainty-based filtering method discussed in Chapter 3. As shown in Section 3.5, the filtering method's effectiveness deteriorates when the unfair testimonies' qualities very close to, or even the same as, the truster's current belief about the trustee. In particular, its effectiveness deteriorates significantly in two cases:

- The provider's willingness is low, i.e. its willingness is fluctuating around $IL = 0.1$, while there are more than 50% *ballot-stuffing* unfair testimonies and p_{unfair} is high, e.g. $p_{unfair} \geq 0.5$.
- And the provider's willingness is high, i.e. its willingness is fluctuating around $IL = 0.9$, while there are more than 50% *badmouthing* unfair testimonies and p_{unfair} is high, e.g. $p_{unfair} \geq 0.5$.

Results shown in Figure 4.8 and Figure 4.9 present the comparison between the credibility model and the filtering method in various experimental configurations.

In Figure 4.8 and Figure 4.9, the filtering method's effectiveness (in terms of MAE) are plotted as lines with labels suffixed by "(O)", while the credibility model's are with labels suffixed by "(C)". It is observed that, the proposed credibility model manages to achieve lower MAE (i.e.

(a) IL=0.1

(b) IL=0.5

(c) IL=0.9

Figure 4.8: Comparison of the credibility model and the filtering method in the presence of ballot-stuffing testimonies

(a) IL=0.1

(b) IL=0.5

(c) IL=0.9

Figure 4.9: Comparison of the credibility model and the filtering method in the presence of badmouthing testimonies

higher evaluation accuracy) than the filtering method in all the experimental configurations studied. Another obvious improvement of the credibility model over the filtering method is that, the credibility model’s performance does not decrease significantly with the increase of unfair testimonies ratio, especially in the two cases mentioned above (see Figure 4.8(a) and Figure 4.9(c)). This validates that the credibility model does improve the filtering method in mitigating the adverse effect of the unfair testimonies.

Then the performance of the credibility model is studied with a real dataset, and compared with other credibility modeling methods, in particular the method proposed by Yu and Singh.

We first study the two methods’ performance without the presence of unfair testimonies. Table 4.1 shows the average MAE achieved by the two methods. MAE achieved with the proposed credibility model is labeled as *Cred*, while the one achieved with Yu and Singh’s method is labeled as *YS*.

Method	MAE Achieved
<i>Cred</i>	0.9736
<i>YS</i>	1.0035

Table 4.1: Comparison of performance without unfair testimonies

It is observed that when there is no presence of unfair testimonies *Cred* outperforms *YS*, though it is not significant. This is because *YS* applies a strategy that is too sensitive to the presence of unfair testimonies. Every testimony that is not *exactly* same as the trustee’s real rating will be considered as unfair, and the witness’s credibility will be decreased accordingly. However, as there is no agent giving unfair testimonies, such a sensitive strategy would reduce the fair testimonies’ contribution to the evaluations, which compromises its performance.

After studying the performance without the presence of unfair testimonies, we examine the two methods’ effectiveness in the presence of unfair testimonies. Effectiveness of both methods (measured by *POU*) in the presence of different ratios of witness agents giving unfair testimonies are shown in Table 4.2. In this table, “Push” denotes the case in which the unfair testimonies are to push the target trustees’ trustworthiness to the highest rating, i.e. $r_{\text{target}} = 5$, while “Pull” denotes the case in which the unfair testimonies are to pull the target trustees’ trustworthiness down to the lowest rating, i.e. $r_{\text{target}} = 1$.

Unfair testimony ratios	POU (Push)		POU (Pull)	
	Cred	YS	Cred	YS
25%	0.0450	0.0738	0.0538	0.0712
50%	0.1003	0.1461	0.1432	0.2014
75%	0.1376	0.2102	0.3305	0.6509

Table 4.2: POU w.r.t. different ratios of witness giving unfair testimonies

The first observation from Table 4.2 is that $\text{POU} > 0$, which shows that the presence of unfair testimonies does bias the trusters' evaluations toward the target ratings. This clearly exemplifies the adverse effect of the unfair testimonies. Nevertheless, *Cred* still manages to achieve higher effectiveness than *YS* in mitigating the adverse effect of unfair testimonies. One of the reasons leading to this observation is that *YS* aggregates all the testimonies without filtering. Therefore, unfair testimonies still exert certain influence over the trusters' evaluation when they are aggregated, though testimonies are assigned different weights when being aggregated. Another reason is that *Cred* also adjusts each testimony based on the corresponding witness' past testimonies, which makes it possible to make a testimony useful even it is unfair.

4.7 Summary

This chapter revamps the uncertainty-based filtering method proposed in Chapter 3 with a credibility model. The proposed credibility model preserves the merits that the filtering method has. For example, it also does not require additional mechanism and information other than the collected testimonies. Besides preserving the filtering method's merits, the proposed credibility model also improves the filtering method in many aspects and has its own merits:

- The proposed uncertainty-based filtering methods only measures the usefulness of the testimony based on the amount of uncertainty it contains. As discussed in Section 3.5, this is not always sufficient, e.g. it is not sufficient in the case in which the witness somehow manipulates the testimony to be “opposite” to the truster's current belief about the trustee. To solve this problem, the credibility model measures the usefulness of the testimony based on how much the testimonies from the same witness have reduced the truster's uncertainty.

This measurement of usefulness is consistent with the effect of third-party testimonies recognized in traditional research on trust [BK95].

- With the same approach, the truster also evaluates the confidence of its own pre-interaction evaluation of direct trust. With the witnesses' credibility and its own confidence evaluated, the credibility model improves the uncertainty-based filtering method with an adaptive filtering criterion, which filters the testimonies that are less useful than the truster's own evaluation of direct trust.
- To facilitate the measurement of the witness's credibility, the credibility model maintains the witness's profile by storing its past testimonies and the truster's own evaluation of direct trust on the same trustees in a contingency table. The information recorded in the profile also enables the testimony adjustment, which helps to further reduce the adverse effect of unfair testimonies. Compared with the testimony adjustment in the trust model proposed by Abdul-Rahman and Hailes [ARH00], in which only the witness's last testimony counts, the testimony adjustment in the proposed credibility model takes into account all the past testimonies of the witness. This treatment is able to capture the difference between the views of the truster and the witness more accurately.
- Another merit of the credibility model is that it is a generic solution to the problem caused by unfair testimonies and can be applied to most computational trust models. It does not require a specific trust model to be used due to its decoupling from the trust model. It can be applied to any computational trust model where outcomes of trust evaluations can be represented in the form of discrete ratings. Since the credibility metric takes discrete ratings as input, it is readily applicable to computational trust models that use discrete ratings to represent the trust evaluations outcomes (i.e. testimonies), e.g. the trust model proposed by Abdul-Rahman and Hailes in [ARH00]. With the preprocessing of testimonies (see Section 4.3.1), the credibility model is also applicable to those trust models which represent trust evaluations outcomes as continuous values, e.g. the basic trust model presented in Section 3.2 (i.e. Beta Reputation System [JI02]).

Experimental results have also been presented to show the effectiveness of the credibility model.

Having presented the proposed credibility model and discussed its advantages, the next chapter presents a more intensive study of its effectiveness in a testbed that simulates the interactions among agents.

Chapter 5

Empirical Evaluation of the Credibility Model

A credibility model was proposed in Chapter 4 to enhance the uncertainty-based filtering method. This chapter conducts a more intensive empirical study of the credibility model's effectiveness. It describes the design of the testbed used to carry out the empirical evaluation. Following that, the methodology used and results achieved are presented. This chapter also presents the results of empirical evaluation with the `Agent Reputation and Trust` testbed, which is a publicly available testbed and provides a common platform for fair comparative study of the proposed credibility model's effectiveness against related work.

5.1 Introduction

The experiments in Chapter 3 are conducted in a way that the consumers frequently interact with same provider. By doing so, it facilitates an intensive study of the adverse effect of unfair testimonies and the proposed filtering method's effectiveness. However, this may not be the case in realistic settings, where there is usually more than one service provider, and an agent usually has few interactions with the same interaction partner [RZ02]. Chapter 4 addresses this problem by conducting experiments with a real dataset. Nevertheless, the experiments in previous chapters primarily concerns about how the accuracy of the trust evaluation is influenced by presence of unfair testimonies. However, users of the trust model usually are concerned not only about the accuracy of the evaluations, but also about whether the model would be beneficial to the interac-

tions with their partners. Therefore, it is desirable that the model can be studied in a configuration that is similar to realistic settings, and that the effectiveness could be evaluated more practically.

Given this, a testbed, named “TrustBed”, is designed to empirically study the effectiveness of the proposed credibility model. This chapter also presents the results obtained in ART (Agent Reputation and Trust) Testbed, the only publicly available testbed.

This chapter is organized as follows. Section 5.2 presents the design of TrustBed. Section 5.3 describes the experimental methodology used. Section 5.4 summarizes the parameters used in the testbed. The experiments conducted with TrustBed are presented in Section 5.5, 5.6, and 5.7. The results obtained with the ART testbed are shown in Section 5.8. A summary of this chapter is given in Section 5.9.

5.2 Design of TrustBed

In TrustBed, there is a number of agents that are assigned a number of problems to be solved. There is more than one agent providing the service that can be utilized by others. Those service-providing agents have differing behaviors in providing services. TrustBed applies the basic trust model (see Section 3.2) to evaluate trustee’s trustworthiness based on the trustee’s past behavior. The proposed credibility model is employed to combat unfair testimonies.

Instead of merely concerning about the accuracy of trust evaluation, the effectiveness of the model is assessed by studying how much benefit the model could bring to its users, namely the agents TrustBed. To achieve this goal, a cost and gain is assigned to each agent for each invoked service usage. Additionally, we exclude all the other factors that can affect an agent’s cost and gain, such as the negotiation of the payment between agents. Thus, the credibility model becomes the only factor that influences agents’ cost and gain. By doing so, the effectiveness of the model can be measured in terms of agents’ cost and gain.

Ideally, the effectiveness should be evaluated in all possible situations. However, the scope for all possible situations would be prohibitively large, making it impractical to study the credibility model in all situations. To address this problem, we apply settings that are believed to be typical for some factors, while for some other factors, we introduce randomness to simulate the real

environment. In doing so, the credibility model can be reasonably evaluated in a wide range of environments.

5.2.1 Interactions among Agents

Similar as in Section 3.4, it is assumed that a service-providing agent does not use services provided by other service-providing agents¹. Thus, the agents in `TrustBed` are separated into two sets: a set of service-providing agents, called *service providers*, and a set of service-consuming agents, called *service consumers*. Usually, the performance of a provider in a particular service is independent from that in another service. Hence, without loss of generality, it is also assumed that provider agents provide the same service.

Each consumer agent is assigned N_m problems. It is assumed that each consumer agent will see the same list of service providers, which contains all the providers in `TrustBed`. As rational agents, consumers will select the providers with higher trustworthiness than those with lower ones.

There is a service-level agreement between the consumer and the selected provider, which specifies the provider's obligations. The outcome of the interaction is binary. The outcome is *successful* if the provider fulfills the obligations in the agreement and the consumer solves the problem with the service provided by that provider². On the other hand, the outcome is considered as *unsuccessful* if the provider defaults its obligations. If the outcome of the interaction is successful, the consumer agent proceeds to the next problem. Otherwise, the consumer will keep trying other providers according to its trust evaluations until it finds a provider to solve the problem. In the worst case, all the providers have been found to default their obligations. The consumer will abandon the problem in this case. Table 5.1 summarizes the agents involved and their interactions in `TrustBed`.

¹If we loose this assumption and service providers are allowed to interact with one another, each service provider will also act as consumer and request testimonies regarding the other providers' behaviors. It incurs extra communication overhead as there will be more messages exchanged among agents. Nevertheless, it does not impact the complexity of the proposed credibility model itself.

²Since the goal of the experiments is to investigate the effectiveness of the proposed credibility model but not how the problems are solved, it is assumed that a problem is solved if and only if the selected provider fulfills its obligations.

Categories of Agent	Primary Attributes	Interactions in TrustBed
Provider	Level of willingness ^a	1. Providing services to consumers
		2. Colluding with consumers to manipulate other consumers' evaluations ^b
Consumer	1. Account Balance ^c	1. Utilizing providers' services
	2. Cost for each service invoked ^c	2. Asking for testimonies from other consumers ^b
	3. Gain if a problem is solved ^c by the selected provider	

^aThis is discussed in Section 5.2.2.

^bThis is discussed in Section 5.2.4.

^cThis is discussed in Section 5.2.3.

Table 5.1: Interaction among Agents in TrustBed

5.2.2 Provider Population Configuration

The behavior of the provider is again controlled by its *willingness*. Providers with various levels of willingness are implemented to simulate the scenario where agents in MAS normally possess differing natures. For example, some are highly willing to fulfill the obligations, some are moderately willing to do so, and some are reluctant to do so. There are four types of providers in TrustBed: *excellent* providers, *good* providers, *ordinary* providers, and *bad* providers. The willingness of the 4 different types of providers are given in Table 5.2. For example, the willingness of the *excellent* providers is 0.8. This means that the *excellent* providers would fulfill their obligations in 8 out of 10 interactions with the consumers.

Category	willingness
Excellent	0.8
Good	0.6
Ordinary	0.4
Bad	0.1

Table 5.2: Configurations of providers' willingness

Moreover, in MAS, it is usually the case that agents' behaviors are not static. Instead, their behaviors change from one interaction to another. To simulate such dynamism, TrustBed also implements the fluctuation of the provider's willingness, similar to what we have done in Section 3.4. An *initial willingness* is assigned to each provider when it enters the testbed. Provider's

initial willingness is set according to the nature of the provider based on Table 5.2. Then in each subsequent interaction with the consumers, the provider can change its willingness. Each change is randomly chosen from the range of $[0, 0.01]$.³

The four types of providers have different ratios of presence in TrustBed. The population mix used is shown in Table 5.3. Such a population can be considered as hostile to the consumer

Category	percentage among all the providers
Excellent	10%
Good	10%
Ordinary	40%
Bad	40%

Table 5.3: Population of the providers

agents because (1) only a small percentage of providers are highly willing to fulfill their obligations, (2) most of the providers only fulfill the obligations occasionally, and (3) a large percentage of providers tend to not fulfil their obligations. This population setting will be used throughout the experiments in this chapter, as well as those in the next chapter. TrustBed chooses such a population as the default setting due to the consideration that if the credibility model is able to show its effectiveness in such a hostile environment, its effectiveness can generally be guaranteed in other settings which are relatively more friendly to the consumers⁴.

5.2.3 Consumer's Cost and Gain

For a consumer agent, say a_s , there is a cost associated with each interaction (i.e. each invocation of provider a_o), which is calculated as $C * TA_{s,o}$ ⁵. $TA_{s,o}$ is the agent a_s 's evaluation of overall trust that it should place in provider a_o . C is a constant which denotes the cost of the consumer if a_o is evaluated to have full trustworthiness (i.e. $TA_{s,o} = 1$). There is also a gain assigned to each consumer agent if the problem is solved by the selected provider, and it is calculated as $G * L_o$.

³Compared with the fluctuation of willingness in the experiments in Chapter 3, the fluctuation of willingness in TrustBed is much smaller. This is to make sure that the behaviors of the four types of providers are still differentiable after the introduction of willingness fluctuation.

⁴We have also run experiments with different population mixtures. Similar trends as shown in this chapter are observed.

⁵ $TA_{s,o}$ is applied here to differentiate the consumer's payment according to its evaluation of the chosen provider. By doing so, the consumer is penalized if the chosen provider's behavior is not as good as its evaluation.

L_o denotes the willingness of the provider when it solves the problem. G is a constant which denotes the gain if the problem is solved by a provider with full willingness (i.e. $L_o = 1$)⁶. With such a strategy of calculation, an agent's cost and gain in different interactions varies according to the selected provider's trustworthiness and willingness.

G and C are set as 5 and 1 respectively, and each consumer agent is allocated an initial balance of $G_{init} = 50$. With such a setting of G and C , each consumer agent can gain some profit if it manages to find a provider with higher willingness to solve the problem with fewer tries. Otherwise, its profit gained from solving the problem decreases with more tries. The consumer may even lose its profit due to too many tries.

5.2.4 Presence of Unfair Testimonies

As discussed earlier in Chapter 1, unfair testimonies can happen because of the differing views of the witness and the truster. This is implemented as *individual unfair testimonies* in TrustBed.

Individual unfair testimonies mean that providers do not collude with consumers, and each consumer gives unfair testimonies independently. In more detail, we simulate five categories of consumer agents with different strategies in giving testimonies: H, P, HP, N, and HN. Consumer agents of category H always reveal their own evaluations of direct trust on the providers honestly as testimonies. Agents in the other 4 categories give different types of unfair testimonies with different magnitudes. Agents in category P and HP give **ballot-stuffing** unfair testimonies, i.e. unfairly positive testimonies that are higher than their real evaluations of direct trust on the providers, while agents in category N and HN give **badmouthing** unfair testimonies, i.e. unfairly negative testimonies that are lower than their real evaluations of direct trust on the providers.

Since the testimonies are given in the form of single scalar rating value, the strategy to simulate the unfair testimonies used in Chapter 3 is not applicable here. Instead, we apply a strategy to generate the unfair testimonies by adding offset to the witnesses' real evaluations of direct trust on the provider. For **ballot-stuffing** unfair testimony, the offset is positive, while the offset

⁶Here, L_o is applied to differentiate the consumer's gain according to its selection of service provider. If the consumer selects a provider with higher willingness, it will receive more gain, and vice versa. By doing so, together with the application of $TA_{s,o}$ in calculating the consumer's cost, the "quality" of the consumer's trust evaluations can be reflected by the gain it has received.

is negative for **badmouthing** unfair testimony. The difference between category P and HP, and N and HN lies in the magnitudes of offset for respective categories. The magnitude of offset for category P and N is randomly chosen from the range of (0.1, 0.4), i.e. the agents of these two categories are giving moderately unfair testimonies. The respective range for category HP and HN is (0.8, 1), i.e. the agents of these two categories are giving highly unfair testimonies.

There are various configurations of unfair testimonies simulated in the `TrustBed`. Each configuration of unfair testimonies is a mix of consumer agents of H category, and agents either giving ballot-stuffing unfair testimonies or badmouthing unfair testimonies. That is, a configuration of unfair testimonies contains a mix of consumer agents of the following category combinations : $\{H, P, HP\}$, or $\{H, N, HN\}$. In each configuration, the agents of category P and HP, or N and HN have equal presence. Configurations differentiate from each other in the ratios of consumer agents giving unfair testimonies. For the convenience of referring to various configurations of unfair testimonies, each configuration is named with the direction of the unfair testimony (i.e. `Pos` for ballot-stuffing and `Neg` for badmouthing) and the ratio of agents giving unfair testimonies. For example, `Neg60` denotes the configuration in which 30% of the consumers are of category N (i.e. giving moderate badmouthing testimonies), 30% are of the category HN (i.e. giving highly badmouthing testimonies), and the rest 40% are of category H. There is a special configuration which is named `Hon`, where all the consumer agents are of category H.

There is another cause of the unfair testimonies. That is, some agents may deliberately give unfair testimonies on certain trustees when being asked by other agents. In this case, the unfair testimonies are usually given with a clear target in order to manipulate the truster agents' evaluations of trust on those target trustee agents. This is implemented as another type of unfair testimonies in `TrustBed`, i.e. the *collusive unfair testimonies*. Collusive unfair testimonies mean that a number of providers colludes with consumers, who intentionally give testimonies in favor of the colluding providers in order to promote the colluding providers' trustworthiness in the eyes of the non-colluding consumers.

To simulate the collusive unfair testimonies, `Bad` providers (see Table 5.2 for the definition of `Bad` provider) are implemented as colluders with a number of consumers. Since the `Bad`

providers generally have the lowest trustworthiness in the eyes of the consumer agents (assuming the evaluations are not manipulated by the presence of unfair testimonies), they can only attract a few consumers to use their services. Hence, they have the motivation to manipulate the non-colluding consumers' evaluations of their trustworthiness. The configuration of the collusive unfair testimonies is similar to the case of individual unfair testimonies. But in the case of collusive unfair testimonies, the testimonies given by the colluding consumers are always ballot-stuffing since the goal is to promote the non-colluding consumers' evaluations of trust on the colluding provider. That is, the agents giving unfair testimonies are only from the category combination of $\{H, P, HP\}$.

5.3 Methodology

5.3.1 Comparison with Related Work

To compare the effectiveness of the proposed credibility model with that of other approaches, the consumer agents in `TrustBed` are separated into three groups. The number of consumer agents in each group is the same and it is denoted as N_c . Consumers in one of the groups are equipped with the proposed credibility model, while the other two groups consist of (1) agents with no mechanism to tackle the presence of unfair testimonies, and (2) agents with other methods to tackle the presence of unfair testimonies. This facilitates a perceptible and objective study of the effectiveness of the proposed credibility model against other methods as they act in identical experimental settings.

For the third group, the method proposed by Yu and Singh⁷ [YS03] is chosen. The reasons for the choice are: first, its applicability has already been validated by some successful applications, such as the P2P system in [YSS04]; second, other than this method, most of the notable methods require additional mechanisms or knowledge, such as the one applied in `ReGreT` [SM03] and the trust model proposed by Jurca and Faltings [JF03, JF04].

Other than the three groups mentioned above, there is another group of consumers, who do not aggregate the third-party testimonies and make the trust evaluations solely based on their own per-

⁷See Appendix A for details about this method.

sonal interaction experiences with the providers. This group is denoted as NoAggr. This group is introduced to investigate whether the aggregation of testimonies improves the consumers' performance in identifying the right providers (i.e. with higher willingness). The aforementioned four groups are labeled as Cred, NoCred, YS, and NoAggr respectively.

5.3.2 Measurement Metrics

The effectiveness of the credibility model is measured in terms of the benefit agents may gain by using it. In the context of TrustBed, the trust model is used to help agents identify the right service providers to maintain their profit. Each individual agent in the four groups uses the same trust model, namely, the basic trust model presented in Section 3.2. The only difference is that each group employs a different model or strategy to combat the unfair testimonies in order to maintain the basic trust model's effectiveness in the presence of unfair testimonies. The effectiveness of the trust model is evaluated by the profit that the agents gain. Moreover, TrustBed excludes other factors that can influence an agent's profit. Therefore, the profit gained by agents in different groups can be interpreted as the effectiveness of the corresponding strategies in addressing the unfair testimonies. Consequently, we record the gain of each agent for each problem. The average over all agents' gains in each individual group is interpreted as the performance of each group. This performance reflects the effectiveness of the strategy used in the corresponding group to mitigate the influence of unfair testimonies. This measurement metric is termed the *gain* metric. The higher the *gain*, the more effective the corresponding strategy is. Since agents in group NoCred employ no mechanism to combat the unfair testimonies, the gain of group NoCred is considered as the baseline. The difference between the gain of group Cred and that of group NoCred as well as group YS will reflect the value of the proposed credibility model.

Besides this *gain* metric, we also measure the *collusion power* in the evaluation of the effectiveness of different strategies in the presence of collusive unfair testimonies. The collusion power is calculated as follows:

$$collusion\ power = \frac{\sum_{a_s \in \mathcal{A}_{nc}} \#try(a_s)}{|\mathcal{A}_{nc}| * N_m} \quad (5.1)$$

where N_m is the number of problems each consumer is assigned to solve, $\mathcal{A}_{nc}$ denotes the set of non-colluding consumers, $|\mathcal{A}_{nc}|$ is the number of consumers in the set of $\mathcal{A}_{nc}$, a_s is a consumer in this set, and $\#try(a_s)$ denotes the number of times consumer a_s has tried to invoke any colluding Bad provider.

In other words, the metric of *collusion power* basically measures the ratio of attempts that the non-colluding consumers have engaged with the colluding Bad providers to the total number of problems that the non-colluding consumers need to solve. If the adverse effect of the collusive unfair testimonies is not fairly addressed, the collusion is able to manipulate the non-colluding consumers' evaluations of trust on the colluding providers, and attracts more non-colluding consumers to use their services, which leads to a higher *collusion power*. Hence, a low value of *collusion power* is expected to show the effectiveness of the proposed credibility model.

For example, suppose there are 20 non-colluding consumers (i.e. $|\mathcal{A}_{nc}| = 20$), and each of them needs to solve $N_m = 100$ problems. If these 20 non-colluding consumers have attempted to invoke the services provided by the colluding providers 4000 times, the *collusion power* in this case would be $\frac{4000}{20 \times 100} = 2$. This value can be interpreted that each of the non-colluding consumers has tried the colluding Bad providers twice on average for every problem.

5.4 Experimental Setup

In summary, the experiments are conducted to study the influence of the unfair testimonies in various configurations. More importantly, we also examine the proposed credibility model's effectiveness in mitigating the adverse effect of unfair testimonies, and the effectiveness of the proposed credibility model is evaluated empirically by the model users' final gain. If the proposed credibility model is effective as expected, the consumer agents should be able to choose provider to complete their problems with few tries even in the presence of unfair testimonies, thus accumulating higher gains. Therefore, a higher gain of the consumer agent (in group Cred) is expected to validate the proposed credibility model's effectiveness.

The experiments are controlled by the following factors:

- The population of the testbed. In each experiment, `TrustBed` is populated with $N_p = 10$ providers. There are different groups of consumers, each of which applies different methods to address the adverse influence of the unfair testimonies. Each group contains $N_c = 100$ consumers.
- The number of problems each consumer needs to solve is denoted as N_m . Each consumer agent is assigned 200 problems, i.e. $N_m = 200$.
- The initial balance of each consumer agent, G_{init} , is set as 50^8 .
- The constant C used to calculate the cost individual consumer needs to pay for each service usage is set as 1.
- The constant G used to calculate the gain of consumer if a problem is solved by the provider is set as $G = 5$.
- The providers' willingness. It is set according to Table 5.2 and Table 5.3.
- The probability that each provider changes its willingness in each interaction, i.e. p_{fluc} . It is set as $p_{fluc} = 1/3$.

Consumer agents in different groups are equipped with the basic trust model presented in Section 3.2 to select the provider. The basic trust model (i.e. Beta Reputation System (BRS)) is mainly controlled by the decaying factor, i.e. the value of λ in Eq. (3.4), and the number of ratings that are taken into account when generating the rating summary, i.e. W in Eq. (3.4). These two parameters are set as $\lambda = 0.9$ and $W = 10$ as in the experiments in Chapter 3. All the consumers in `TrustBed` use these same settings.

Table 5.4 summarizes the parameters used in the experiments.

There is also a number of variables, which control the presence of unfair testimonies in `TrustBed`. In the experiments, we apply different settings of these variables to generate various

⁸Setting $G_{init} = 50$ is to make sure each consumer agent has enough balance to pay for the service and testimonies it requested initially. Changing this setting does not impact the results presented in the following sections, as the results study the average gain of each consumer agent instead of the balance.

Parameter	Description	Value
N_p	number of providers	10
N_c	number of consumers in each group	100
N_m	number of problems assigned to each consumer	200
G_{init}	initial balance of consumer agent	50
C	The constant used to calculate consumer agent's cost for each service usage	1
G	The constant used to calculate consumer agent's gain for each solved problem	5
	providers' willingness	set according to Table 5.2 and 5.3
p_{fluc}	The probability that each provider changes its willingness	1/3
λ	decaying factor in the basic trust model	0.9
W	number of past ratings taken into account in the basic trust model	10

Table 5.4: Experiment parameters

configurations of unfair testimonies, and to study the effectiveness of the credibility model in a wide range of settings. Five individual experiments are run with each setting of these variables, and the average of the aforementioned metrics obtained in the five experiments are taken as the final result of the corresponding configuration.

The proposed credibility model itself also has a parameter for fine tuning. This is the scale of discrete ratings that represent the discretized testimonies, namely, Z in Figure 4.1. The default value of Z is set as 10. We also investigate the impact of the value of Z over the effectiveness of the proposed credibility model by running experiments with different values of Z .

5.5 Effectiveness in the Presence of Individual Unfair Testimonies

First of all, as the baseline, we study the performance of each group with the unfair testimonies configuration Hon, where none of the agents give unfair testimonies. Table 5.5 summarizes the

average gain of each group. It is observed that group NoAggr achieves the lowest performance (in terms of the *gain* metric). Figure 5.1 plots the average gain of the agents for each problem in each individual group. It is observed from Figure 5.1 that for almost every problem, group NoAggr is outperformed by the other three groups which aggregate the testimonies. This observation shows that solely depending on an agent's personal interaction experiences with the provider is not sufficient to make a trust evaluation that truly reflects the provider's behavior. This observation also confirms the effect of third-party testimonies [BK95], that aggregation of third-party testimonies does reduce the consumer agent's uncertainty about the provider's behavior.

Group	Average Gain
Cred	3.45
NoAggr	2.66
NoCred	3.09
YS	2.95

Table 5.5: Comparison of performance without unfair testimonies

Figure 5.1: Average gain of each group without unfair testimonies

We then run the experiments with different configurations of unfair testimonies to see whether each group is able to maintain its performance in the presence of unfair testimonies as in the case where there were no unfair testimonies.

Table 5.6 lists the average gain of each group, and Figure 5.2 plots the average gain of the agents in each individual group for each problem in the presence of unfair testimonies Pos100.

Group	Average Gain
Cred	3.45
NoAggr	2.64
NoCred	2.37
YS	2.71

Table 5.6: Performance of each group in the presence of individual unfair testimonies Pos100

Figure 5.2: Average gain of each group in the presence of individual unfair testimonies Pos100

In the presence of Pos100 unfair testimonies, all the consumer agents are giving **ballot-stuffing** unfair testimonies (50% gives moderately unfair testimonies, and 50% gives highly unfair testimonies). In this case, group NoAggr manages to maintain its performance as in the case of no unfair testimonies. This is because it is not influenced by the presence of unfair testimonies since it does not aggregate testimonies at all. However, group NoCred's performance has been greatly aggravated by the presence of unfair testimonies, and it is outperformed by the other groups in almost every problem as shown by Figure 5.2. This is due to the fact that it blindly aggregates all the testimonies, including the unfair testimonies. Compared with group NoCred, group YS manages to mitigate the adverse influence of the unfair testimonies, but the performance is not as good as in the case of Hon, i.e. no unfair testimonies. Its performance is below that of group Cred, which is able to maintain its performance even in such an extreme case at a comparable level as in the case of Hon.

Experiments have also been conducted to study the four groups' performance in presence of other configurations of unfair testimonies, namely, Pos80, Pos60, Pos40, Pos20, Neg20, Neg40, Neg60, Neg80, Neg100. Due to the large number of experiments conducted, a sum-

mary of the average gain over all the problems of the four groups are presented in Figure 5.3. Group NoAggr's performance is relatively stable for all configurations of unfair testimonies

Figure 5.3: Performance of different groups in the presence of various configurations of individual unfair testimonies

considered. As discussed above, this is because it is not influenced by the presence of unfair testimonies as it does not aggregate testimonies at all. However, since the provider's behavior is changing from one interaction to another, solely depending on the consumer agent's personal interaction experience with the provider is not sufficient for the consumer agent to make a trust evaluation that can truly reflect the provider's behavior. Hence, group NoAggr achieves a performance lower than the other groups that aggregate the testimonies (i.e. group Cred and group YS, but not NoCred), though its performance is stable in the presence of unfair testimonies.

Group NoCred's performance exhibits a great fluctuation in the presence of various configurations of unfair testimonies. Generally, group NoCred's performance (in terms of average gain) decreases when more consumers giving unfair testimonies. This is because it aggregates all the testimonies blindly. The presence of unfair testimonies exerts a great impact over the consumer agents' evaluations of trust on the providers, which confuse the consumer agents' identification of the right providers (i.e. those with higher willingness to fulfill their obligations). Consequently, the agents in group NoCred need more tries before each of their problems can be solved, leading to their lower gain for each problem. This observation exemplifies the adverse effect of unfair testimonies.

Both group YS and group Cred manage to reduce the adverse effect of unfair testimonies, and thus their performance is more stable, as compared with group NoCred. Nevertheless,

group YS's performance is still lower than that of group Cred. Group YS is even outperformed by group NoCred in the case of Hon. This is because, group YS applies a strategy that is too sensitive to the presence of unfair testimonies. That is, every testimony that is not *exactly* the same as the truster's post-interaction direct trust on the trustee will be considered as unfair, and the witness' credibility will be decreased accordingly. However, as there is no agent giving unfair testimonies in the case of Hon, such a sensitive strategy would reduce the fair testimonies' contribution to the trust evaluations, making it even less effective than the strategy of not measuring the credibility (i.e. as group NoCred does).

As expected, group Cred consistently outperforms group NoAggr and NoCred. Moreover, group Cred also outperforms group YS. Several reasons contribute to group Cred's superiority over group YS:

- A consumer in group YS first aggregates the testimonies as a weighted mean of all the available testimonies. Then the aggregation of testimonies is linearly combined with the consumer's own evaluation of direct trust to obtain the overall trust on the provider. The factor assigned to the aggregation of testimonies is based on the number of past interactions with the provider. That is, after the consumer has a number of interactions with the provider (the number equals to or is more than the value of W , see Appendix A for details), the testimonies are not aggregated. However, since the behavior of the provider changes from one interaction to another, not taking account of the third-party testimonies after a number of interactions with the providers will deteriorate the trust evaluation.

In contrast, consumer agents in group Cred apply a more adaptive strategy in aggregating the testimonies. Only those more *useful* than its own evaluation of direct trust on the provider will be aggregated.

- Consumers in group YS aggregate all the testimonies without filtering. Although different testimonies are assigned different weights when aggregating the testimonies, unfair testimonies still exert certain influence over the consumer agents' trust evaluation when they are aggregated.

- Apart from only filtering the testimonies based on corresponding witnesses' credibility, consumer agents in group `Cred` also adjust each testimony based on the corresponding witness' past testimony-reporting. This enables group `Cred` to maintain its performance even in the extreme case as `Pos100` where all the consumers give unfair testimonies, since the testimony adjustment helps to make a testimony "useful" even it is unfair.

5.6 Effectiveness in the Presence of Collusive Unfair Testimonies

We also study the performance of different groups in the presence of collusive unfair testimonies. The collusive unfair testimonies are implemented as a number of consumers colluding with `Bad` providers. Those colluding consumers give **ballot-stuffing** unfair testimonies to manipulate non-colluding consumers' evaluations of trust on the colluding `Bad` providers. As the colluding consumers are always giving ballot-stuffing unfair testimonies, the configurations of unfair testimonies considered are `Hon`, `Pos20`, `Pos40`, `Pos60`, `Pos80`⁹.

For the measurement of the effectiveness, the *gain* metric is measured again, as in the case of individual unfair testimonies. However, this metric is only calculated over the non-colluding consumers, since the colluding consumers' unfair testimonies only affect the non-colluding consumers' evaluations of trust on the colluding `Bad` providers. Beside the *gain* metric, we also measure the *collusion power*, which reflects whether the colluding `Bad` providers have succeeded in manipulating the non-colluding consumers' evaluations of trust. Figure 5.4 presents the average gain and collusion power of each group in the presence of various configurations of collusive unfair testimonies. Since the performance of group `NoAggr` is not influenced by the presence of unfair testimonies, it is not included in the figures for the sake of simplicity.

The first observation from Figure 5.4 is that collusive unfair testimonies does lure the non-colluding consumers to try the colluding providers more frequently. In Figure. 5.3, group `NoCred` was still able to achieve an average gain above 2.3 even in the presence of most powerful individual unfair testimonies, i.e. `Pos100` and `Neg100`. However, as shown in Figure. 5.4 (a), presence of collusive unfair testimonies `Pos20` already pulls group `NoCred`'s average gain down to 2.3.

⁹`Pos100` is not considered since there is no non-colluding consumer in this case.

(a)

(b)

Figure 5.4: Performance of different groups in the presence of various configurations of collusive unfair testimonies: (a) gain; (b) collusion power

This observation shows that the adverse effect of collusive unfair testimonies is more powerful than that of individual unfair testimonies in influencing the trust evaluations.

The proposed credibility model does mitigate the adverse effect of collusive unfair testimonies, as the *collusion power* of group Cred is kept below 0.25 in the presence of various configurations of collusive unfair testimonies. However, the average gain of group Cred is still lower than that in the presence of individual unfair testimonies. This also shows that the collusive unfair testimonies are more powerful in influencing group Cred than the individual unfair

testimonies. This is because, the colluding consumers do not give unfair testimonies on non-colluding providers. Their testimonies on other providers are still considered as useful for the non-colluding consumers. Hence, the colluding consumers are able to “cheat” the credibility model and achieve a high credibility and then give unfair testimonies on the colluding providers. This makes the non-colluding consumers’ filtering and adjustment of the testimonies inaccurate, which consequently lowers their average gain. Nevertheless, due to the reasons discussed at the end of Section 5.5, group Cred still outperforms group YS.

5.7 Influence of Value of Z

So far, the proposed credibility model applied by consumers in group Cred uses a Z value of 10. To study the influence of this parameter on the performance of the proposed credibility model, we change the value of Z and re-run the experiments. It is noted that the experiments are re-run with only the presence of collusive unfair testimonies due to the consideration that the credibility model is less successful in this situation. The smallest possible value of Z is 2, that is, the continuous testimonies are pre-processed to become binary ratings. We have also studied the other possible values of Z , i.e. $Z = 5$ and $Z = 7$.

Figure 5.5 presents the collusion power obtained with different settings of Z value. It is observed that the effectiveness of the credibility model (in terms of collusion power) is generally improved with an increase of Z value. This is because this parameter basically controls the credibility model’s “sensitivity” to the witness’ unfair testimonies. With a smaller value of Z , the discretization of the testimonies is coarser, which makes the discrimination between fair and unfair testimonies relatively coarser too. Hence, a witness is able to achieve a high evaluation of credibility even if its testimonies are greatly different from the truster’s own post-interaction direct trust. This will consequently decrease the effectiveness of the testimony filtering and adjustment. On the other hand, with a larger value of Z , the discrimination between fair and unfair testimonies becomes relatively finer. The evaluation of the witness’ credibility can thus reflect the usefulness of the witness’ past testimonies, which consequently enhances the effectiveness of the filtering and adjustment.

There is a trade-off in the selection of Z value. It is expected that the effectiveness of the proposed credibility model would be enhanced with the larger value of Z , i.e. $Z \geq 10$. Nevertheless, the value of Z determines the size of the profile that the truster maintains for each witness. The storage for maintaining the profiles increases with Z . Based on the results shown in Figure 5.5, an optimum value of Z is in the range of $[5, 10]$ as this achieves a balance of effectiveness and storage overhead.

Figure 5.5: Influence of parameter Z

5.8 Empirical Evaluation with the ART Testbed

The effectiveness of the proposed credibility model has also been evaluated by the Agent Reputation and Trust (ART) testbed¹⁰, which is a publicly available testbed. There is also an annual competition based on the ART testbed held by the *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, which provides a common platform for fair comparative study of different credibility models. An agent equipped with our proposed credibility model took part in the competition held in 2006, and was ranked 2nd among all the participants. This has validated the proposed credibility model's effectiveness in combating the unfair testimonies in a fair and objective manner. This result is presented in detail in the following subsections.

¹⁰ART testbed: <http://www.lips.utexas.edu/art-testbed/>

5.8.1 Overview

The ART testbed simulates the interactions in art appraisal domain, in which agents function as painting appraisers, each with varying levels of expertise in different artistic eras. Each round of simulation consists of a series of discrete time steps. In each time step, clients buy (with a cost defined as “client fee”) appraisals for paintings from different eras; if the requested appraising agent does not have enough expertise to complete the appraisal, it can also buy (with a cost defined as “opinion cost”, which is much less than “client fee”) opinions from other appraiser agents. The accuracy of each opinion generated by individual appraiser depends on how much (not necessary the same as “opinion cost”) the appraiser pays in appraising the painting. There is no guarantee about the fairness of the replies from other appraisers. That is, an appraiser agent may provide unfair testimonies regarding the values of paintings. Hence, before the appraiser agent requests for opinions, it evaluates the reputation of other appraiser agents, and requests opinions from more reputable agents. In the context of ART testbed, an agent’s *reputation* basically captures its *credibility* in giving opinions of paintings [FKM⁺05,FKM⁺06,OKY06]. To evaluate an agent’s reputation, the appraiser would also buy (with a cost defined as “reputation cost”, which is much less than “opinion cost”) others’ reports on this agent’s reputation. Similarly, the agent evaluates the *sociability* of other appraisers before requesting for reputation reports. An appraiser’s *sociability* denotes the credibility of this appraiser in providing reputation reports [FKM⁺05,FKM⁺06,OKY06]. After the appraiser collects the opinions bought from other appraisers, those opinions are aggregated with its own opinion to generate the final appraisal of the painting, which is finally passed to the requesting client.

At the end of each time step, the true value of the painting is revealed to the appraising agent. This allows the appraising agent to compare its final appraisal with the painting’s true value, which helps to update its evaluations of the other agents’ reputation and sociability consequently. An appraiser agent’s updated evaluations of other appraisers’ reputation and sociability will influence its selections of agents from which to buy the opinions or reputation reports in next time step. At the same time, an agent that provides more accurate appraisal will more likely attract

future clients. Thus, each appraiser's client share in the next time step is determined and clients are assigned to the appraisers before the next time step starts.

The aim of each agent is to achieve a higher bank balance at the end of the simulation. Like the testbed designed in Section 5.2, ART testbed also excludes other factors that can affect the agents' bank balance, such as the negotiations between appraisal agents about the cost of each request. Thus, it makes the model used to evaluate the agents' reputation and sociability the only factor that influences the appraiser agent's bank balance.

5.8.2 Architecture of ART Testbed

The ART testbed consists of a simulation engine and a database. The simulation engine runs the simulation by manipulating the communications among agents through various types of messages and making the necessary computations (e.g. the generation of agents' own opinions and final appraisals of the paintings). The database is used to host information about each round of simulation.

The parameters are set at the beginning of each run of simulation. Some parameters, such as the *expertise* values of each appraiser agent for different eras, are set randomly by the simulation engine. Most of the other parameters have predefined values which remain fixed throughout the execution of each simulation run. These parameters are summarized in Table 5.7, and are revealed to all the agents that participate in the simulation [FKM⁺05,FKM⁺06,OKY06].

The simulation engine enforces the following sequence of actions of each appraiser agent [FKM⁺05,FKM⁺06,OKY06]:

- (1) *prepareReputationRequest()*: The appraiser agent determines which agents to request reputation information about other agents from.
- (2) *prepareReputationAcceptsandDeclines()*: The appraiser agent determines whether to answer the incoming reputation requests.
- (3) *prepareReputationReplies()*: According to the received accept or decline messages, the appraiser agent generates reply messages for the agents that have requested reputation information.

Parameter	Description
Time Steps	number of discrete timesteps in the simulation
Painting-Eras	number of possible eras to which paintings can belong
Client Fee	the fixed fee paid by clients to appraisers for painting appraisals
Opinion Cost	the fixed fee paid by opinion requesters to opinion providers (Opinion Cost $\ll$ Client Fee)
Reputation Cost	the fixed fee paid by reputation requesters to reputation providers (Reputation Cost $\ll$ Opinion Cost)
Sensing Cost Accuracy ratio	relating opinion-generation cost to resulting opinion error distribution standard deviation
Old Client Share Influence	the influence of previous client share on client share in the next timestep
Average clients per Appraiser Agent	Determines the total number of clients (i.e. this parameter times the number of agents in the simulation)

Table 5.7: Parameters of ART testbed and competition

- (4) *prepareOpinionRequests()*: The appraiser agent determines which agents to ask for opinions about the paintings that the clients have requested it to evaluate.
- (5) *prepareOpinionCertainties()*: After receiving the opinion requests, the appraiser agent determines the certainties of its opinions if it wants to respond to the requesting appraisers.
- (6) *prepareOpinionRequestConfirmations()*: According to the certainty values that the responding agents send, the requesting appraiser agent sends confirmations if it really wants to buy opinions from those responding agents.
- (7) *prepareOpinionCreationOrders()*: For each incoming opinion request and the appraiser agent's own appraisal assignments, the appraiser agent orders the simulation engine to generate its own opinions.
- (8) *prepareOpinionProviderWeights()*: For each bought opinion as well as its own opinion, the appraiser agent sends the corresponding weights of each opinion to the simulation engine. As the simulation engine has both the opinion values and the weights, it can make the final calculation of the appraiser's appraisal of the paintings at a later stage.
- (9) *prepareOpinionReplies()*: The appraiser agent responds with opinions on paintings to agents who have bought opinions from it.

Each individual appraiser agent employs a certain strategy to evaluate another agent's reputation and sociability, which is embedded in the above nine actions. The employed strategy also helps the appraiser agent to determine (1) how to buy information (both opinions on paintings and reputation reports on other appraiser agents), and (2) how to reply to others' information requests (requests for opinions on paintings and reputation reports). The communication among the participating agents is handled by the simulation engine via exchange of various types of messages.

In each round of simulation, there is usually more than one participating appraiser agent. They compete with each other for client share in order to achieve a higher bank balance. In ART testbed, the sole factor that influences agents' bank balance is the agents' strategies to model other agents' reputation and sociability. The effectiveness of agents' strategies can thus be measured according to their *bank balance*. Therefore, *bank balance* is employed as the metric to measure the effectiveness of participating agents' strategies [FKM⁺05, FKM⁺06, OKY06].

5.8.3 Results of ART Competition 2006

Table 5.8 lists the setting of the parameters fixed for the final rounds of ART competition 2006.

Parameter	Value
Time Steps	60
Painting-Eras	10
Client Fee	100
Opinion Cost	10
Reputation Cost	1
Sensing Cost Accuracy Ratio	0.5
Old Client Share Influence	0.1
Average clients per Appraiser Agent	20

Table 5.8: Parameter settings of ART competition 2006

Due to the large number of participants in ART competition 2006, this subsection only shows the results of the final rounds of the competition¹¹, in which there were 5 participants, namely iam, Neil, Frost, sabatini, and Joey.

¹¹There were a total of 17 participants, each of which engaged in 6 rounds of simulation with others. Every round of simulation had 5 participants. Complete results of ART competition 2006 is accessible at http://www.lips.utexas.edu/art-testbed/competition_results.htm.

Agent `Neil` was equipped with the credibility model presented in this thesis. Recall that the proposed credibility model is used to measure the usefulness of witness agents' testimonies in reducing the truster agents' uncertainty about the trustee. When it is applied in the context of ART testbed, the trustee is in fact each individual painting, which is not a living object and does not have its own behavior. The trustworthiness of the trustee is basically the true value of the painting, and it is set randomly by the simulation engine, instead of being evaluated with a trust model based on its past behaviors. From agent `Neil`'s viewpoints, the other appraiser agents act as witnesses of the trustees (i.e. the paintings). `Neil` evaluates the witnesses' reputation with the proposed credibility metric (see Eq. (4.2)). Similarly, the proposed credibility metric is applied to evaluate agents' sociability. At the end of each timestep, `Neil` receives information about the true values of the paintings for which the clients have requested his appraisals, and updates its evaluations of other appraiser agents' reputation and sociability according to procedure shown in Figure 4.5.

According to the filtering criterion in the credibility model (see Eq. (4.3)), `Neil` only buys opinions and reputation reports from appraiser agents that have a higher reputation and sociability than its own confidence in appraising the paintings and evaluating other appraisers' reputation. For each incoming opinion request, the cost that `Neil` pays to generate an opinion is calculated as:

$$Rep * cp \tag{5.2}$$

where Rep denotes `Neil`'s evaluation of the opinion requester's reputation, while cp denotes the cost that `Neil` pays to generate an opinion for paintings that the clients have requested it to appraise. That is, `Neil` employs a reciprocal strategy in generating opinions for other appraisers, namely, it generates opinions of various accuracy depending on the requesting appraisers' reputation.

Other than agent `Neil`, only the strategy of agent `Frost` is publicly available, which is reviewed in Section 2.3.3.4. Furthermore, similar to agent `Neil`, `Frost` applies a reciprocal strategy in responding to the incoming opinion requests. Nevertheless, the description of `Frost`'s

strategy in [OKY06] does not specify clearly how it determines the cost it pays to reply incoming opinion requests.

There were a total of 10 rounds of simulation conducted. In each individual round of simulation, every agent's bank balance is recorded. The average of each agent's bank balance in all the 10 simulations is taken as the final score of the corresponding agent.

Table 5.9 summarizes the final results of ART competition 2006. It is seen that agent `Neil`,

Agent	Bank balance (Average)
<code>iam</code>	131583
<code>Neil</code>	103023
<code>Frost</code>	102577
<code>sabatini</code>	101411
<code>Joey</code>	92479

Table 5.9: Results of ART competition 2006

equipped with the proposed credibility model, outperformed almost all the other agents in the competition. This result has validated the proposed credibility model's effectiveness in a fair and objective manner. Agent `Neil` was only outperformed by agent `iam`. However, currently, it is not possible to conduct a detailed comparative study of the strategy used by `iam` and `Neil`, since the former is still not publicly available.

5.9 Summary

In Chapter 4, we have revamped the uncertainty-based filtering method with a credibility model. We have seen the advantages of the proposed credibility model. For example, the measurement of credibility is consistent with the effect of third-party testimonies recognized in traditional research on trust [BK95]; the testimony adjustment is able to capture the difference between the views of the truster and the witness more accurately than other methods, e.g. the one in the trust model proposed by Abdul-Rahman and Hailes [ARH00]; it does not explicitly require a specific computational trust model to be used. All these advantages contribute to make the proposed credibility model a generic solution to address the adverse effect caused by the presence of unfair testimonies.

The empirical evaluation in TrustBed further validates its effectiveness in mitigating the adverse effect of unfair testimonies. A number of experiments has been conducted to investigate the effectiveness of the proposed credibility model in mitigating the adverse effects of various configurations of unfair testimonies. In all the experiments, the proposed credibility model is able to maintain the basic trust model's performance at a comparable level to that without the presence of unfair testimonies. It also consistently outperforms related work, e.g. the method proposed by Yu and Singh [YS03]. The results in the ART competition 2006, which is a common platform for comparative studies of different credibility models, have also validated the proposed credibility model's effectiveness in a fair and objective manner.

With the proposed credibility model incorporated, a trust model works as Figure 4.7 shows. However, in Figure 4.7, there is one issue left unaddressed. That is, how the agent is able to find the relevant witness and collect the testimonies in an efficient manner (i.e. line 2 in Figure 4.7). Related work usually assumes that agent would directly ask all the other agents for testimonies. Such an assumption is unrealistic as there is usually a large number of participating agents. The communication cost would be high if each individual agent asks for testimonies from every other agent directly. In fact, one of the reasons that ART testbed is not employed as the primary empirical evaluation tool in this thesis, though it is publicly available, is that agents in ART testbed always contact each other directly.

To address this issue, we propose a credibility-aware referral process on top of the proposed credibility model in the next chapter.

Chapter 6

Counteracting the Unfair Referrers: Measuring the Referral-Credibility

We have discussed how to mitigate the adverse effect of unfair testimonies in the previous chapters (Chapter 3, 4, and 5). This chapter describes how an agent is able to find relevant witnesses and collect their testimonies. This chapter proposes a credibility-aware referral process on top of the proposed credibility model. It also proposes an approach to tackle the presence of malicious referrers during the referral process, which would intensify the adverse effect of unfair testimonies. Empirical evaluation of the proposal is also presented.

6.1 Introduction

One of the fundamental issues in the research on trust modeling in MAS is how to mitigate the adverse effect of unfair testimonies [JIB07]. In previous chapters, we have presented our effort to address this issue. This chapter describes how an agent is able to find relevant witness and collect their testimonies (i.e. line 2 in Figure 4.7). This is also an important issue in the research on trust modeling in MAS [YS05, YS03, YSS04], since the working of the trust model depends on the availability of the testimonies. As it can be seen from the review in Chapter 2, most existing notable work does not address this problem explicitly, but assumes that the agent would ask for testimonies from all the agents in MAS. However, this assumption is unrealistic as there is usually a large number of agents present in MAS. The communication cost would be prohibitively high if each individual agent asks for testimonies from every other agent. Hence, a

more efficient approach is desired. To address this issue, this chapter proposes a credibility-aware referral process to facilitate agents' testimony discovery on top of the proposed credibility model. The basic idea of the credibility-aware referral process is that each truster agent contacts other agents for testimonies, and requests those agents to collect testimonies for it iteratively if they are not witnesses. And during the referral process, the truster agents adaptively direct the testimony discovery toward those agents with high credibility.

It is not possible that the truster would evaluate the credibility of all the agents in MAS. Hence, the truster agent needs to depend on credibility report from the agent that is currently being contacted. However, depending on credibility reports from other agents induces the presence of *malicious referrers*, i.e. those agents who do not give unfair testimonies themselves, but direct the truster to agents who give unfair testimonies. For example, agent a_s is evaluating the trust it should place in agent a_o . Agent a_s first asks agent a_{w_1} and a_{w_2} for testimonies. Agents a_{w_1} , a_{w_2} , a_{w_3} , and agent a_o have formed a collusion. Every time agent a_{w_1} and a_{w_2} receive a request from other agents outside the collusion for testimonies on agent a_o , they decline to give testimonies and intentionally refer the truster agent a_s to a_{w_3} , who always give unfairly positive testimonies on a_o . The presence of malicious referrers would intensify the adverse effect of unfair testimonies. Despite the fact that it would intensify the adverse effect of unfair testimonies, the presence of the malicious referrers is left unaddressed in the existing work.

Motivated by this problem, apart from presenting the credibility-aware referral process, this chapter also proposes an approach to counteract the presence of malicious referrers. The main idea of this approach is to measure each agent's reliability in reporting other agents' *credibility* during the referral process. Since it is about an agent's reliability in reporting credibility, it is termed as the *referral credibility* of the agent in order to differentiate it from the *credibility* of the agent.

This chapter is organized as follows. First, the credibility-aware referral process is presented in Section 6.2. Section 6.3 discusses how the introduction of referral-credibility modeling revamps the referral process. Following that, the approach to model and update referral-credibility is presented in Section 6.4. We empirically evaluate the effectiveness of the modeling of referral-credibility in Section 6.5. Finally, Section 6.6 summarizes this chapter.

6.2 Credibility-aware Referral Process

As it can be seen from the review in Chapter 2, most existing notable work does not consider how the truster agent is able to find relevant witnesses. There is only one notable work that has addressed this issue, namely, the referral process proposed by Yu and Singh [YS02, YSS04]. In this chapter, we exploit the design of this referral process, and improve it with the credibility-awareness.

The original idea of the referral process is similar to the word-of-mouth process which is commonly seen in human society. That is, each agent maintains direct relationships with a number of agents, and requests each of those agents to find witnesses that the latter have direct relationships with in an iterative manner. In other words, the truster agent a_s finds the relevant witnesses on the trustee agent a_o with the following procedure:

- Agent a_s contacts all those agents with which it has direct relationships. Those agents are called agent a_s 's neighbors. Usually, agent a_s has a small number of neighbors, and the number of neighbors it maintains is denoted as N_h .
- Each of a_s 's neighbors, upon being contacted, tries to recall experiences related to agent a_o 's past behaviors. It becomes a witness of agent a_o , if it does locate such experiences. It will then evaluate its direct trust on agent a_o and return the outcome of evaluation as testimony to a_s .
- If any of agent a_s 's neighbors cannot discover such experiences, it will return *referrals* containing its own neighbors to agent a_s .
- If agent a_s receives referrals from an agent, it will further contact those referrals to check whether they are witnesses. Moreover, truster agent a_s will contact an agent only once during one instance of referral process, even if this agent has been returned more than once as referral.
- The above process is carried out repeatedly on the contacted referral until the length of the *referral chain* reaches the predefined threshold U . Here, a referral chain is made up of the

series of referrals that agent a_s has contacted. Its length is basically the number of agents that a_s has contacted by following that referral chain. Setting a threshold is to prevent a truster from endlessly finding relevant witnesses. This is necessary due to the consideration that the resources available to each agent may be limited (in terms of agent's memory and the communication cost) and, more importantly, the truster usually has limited time for trust evaluation before it has to engage in the interaction.

Figure 6.1: Referral Process

This process can be presented as a sequence diagram as shown in Figure 6.1. In summary, the referral process is basically an iterative process in which each referral is requested by the truster to give testimonies or referrals. Hence, for an agent along the referral chain, it may take on more than one role. Names are assigned to these roles for ease of reference. The agent who returns referrals is called the *referrer*. A referral, when contacted by the truster, can in turn become a *witness* or *referrer* based on whether it has previous experience with the trustee: it is a *witness* if it has interacted with the trustee before; it can also become a *referrer* if it is not a witness and returns referrals. An agent, when taking the role as a referrer, is denoted as agent a_r , while it is denoted as agent a_t when taking the role as a referral.

In the original design of the referral process, all the referrals returned by a referrer will be contacted by the truster agent. However, since there are unfair testimonies, some would be more credible than the others among all the referrals. It is rational that the truster would prefer to contact agents with higher credibility. This inspires the design of the credibility-aware referral process. That is, with the referrals' credibility known, not all the returned referrals will be contacted. Instead, the probability with which a particular referral is contacted by the truster is biased, in the sense that it is proportional to the credibility of that referral.

To facilitate the credibility-awareness, each referrer is also required to report the referral's credibility. The credibility-awareness depends on the referrer's report of the referral's credibility. This is due to the fact that it is not possible for the truster agent to meet and evaluate all the agents in MAS.

It should be noted that the credibility-aware referral process does not assume the transition notion of credibility. For example, agent a_2 is one of agent a_1 's referrals, and a_1 is truster agent a_s 's neighbor. a_1 's credibility is x , while a_2 's is y . If credibility is transitive, the probability of a_2 being visited could naturally be calculated as $x * y$, which is basically a_2 's credibility discounted by a_1 's credibility. The advantage of transitive credibility is that eigenproblem-based method can be applied to calculate the probability that an agent would be visited by the truster agent during the referral process, like the one used in PageRank algorithm [PBMW98]. This works when the referrer is credible enough, which corresponds to the metaphor of "friend's friends should also be friends". Nevertheless, it does not work when the referrer is not credible, i.e. the referrer is "enemy". In this case, there are two possible consequences: (1) enemy's enemies should be friends as they share interest in fighting same opponent, and (2) enemy's referrals should not be taken into account as it has strong incentive to cheat. However, transitive credibility is not able to distinguish these two cases. Given this, the credibility-aware referral process does not assume credibility is transitive. Consequently, eigenproblem-based approach is not applicable here. Instead, the credibility-aware referral process only helps truster agents to discover possible witnesses via referral. Whether to visit an agent is left to the truster agents to decide, which is discussed in the rest of this section.

Suppose the truster agent a_s is currently requesting testimonies from agent a_r . As agent a_r is not a witness, it returns a set of referrals to truster a_s . Among those referrals, the probability that a referral, say agent a_t , is contacted by the truster a_s is determined as follows:

$$\Pr(v_{s,t}) = \frac{TT_{r,t}}{\sum_{a_t \in \mathcal{R}_c} TT_{r,t}} \quad (6.1)$$

where $TT_{r,t}$ is the credibility of agent a_t as reported by agent a_r , $\mathcal{R}_c$ is the set of referrals returned by a_r , and $\sum_{a_t \in \mathcal{R}_c} TT_{r,t}$ is the sum of the credibility of all the referrals returned by agent a_r . Figure 6.2 summarizes the credibility-aware referral process.

Procedure `referral( $s, r$ )`

a_s : the truster agent,

a_o : the trustee agent,

a_r : the agent that is currently being contacted by a_s during the referral process,

$\mathcal{R}$: the set of agents that have been contacted by a_s before,

$\mathcal{R}_c$: the set of referrals returned by agent a_r .

```

1: if  $a_r$  is a witness of  $a_o$  then
2:    $a_r$  returns its evaluation of direct trust on  $a_o$  as testimony to  $a_s$ 
3: else
4:    $a_r$  returns a set of referrals  $\mathcal{R}_c$ 
5: end if
6: for all  $a_t \in \mathcal{R}_c$  do
7:   if  $a_t \in \mathcal{R}$  or  $length(a_t) \geq U$  then
8:      $a_s$  ignores  $a_t$ 
9:   else
10:    with probability  $\frac{TT_{r,t}}{\sum_{a_t \in \mathcal{R}_c} TT_{r,t}}$ ,  $a_s$  contacts  $a_t$ 
11:    add  $a_t$  to  $\mathcal{R}$ 
12:    referral( $s, t$ )
13:   end if
14: end for
    
```

Figure 6.2: Credibility-aware referral process

A simplified example of the referral process is shown below.

Example 6.1 *In the context of Figure 6.3, truster agent a_s is finding witnesses on trustee agent a_o , with threshold on the referral chain length $U = 4$.*

Agents a_1 and a_2 are agent a_s 's neighbors. Both of them are not witnesses of the trustee. Hence, they return referrals to truster agent a_s . Agent a_3 and a_4 are the two referrals returned by

Figure 6.3: Example of referral process

agent a_1 , while agent a_2 returns agent a_5 and a_6 as referrals. Based on the probabilistic strategy of the credibility-aware referral process, a_s does not contact a_3 (the dotted line between a_1 and a_3 in the figure) but only contacts a_4 . With the same probabilistic strategy, agent a_s contacts both agent a_5 and a_6 .

Agent a_4 is not a witness, and returns a_7 and a_8 as referrals. Agent a_6 is a witness (shown as double circle in the figure) of a_o , and returns testimony to truster agent a_s . a_5 is not a witness, and returns a_1 and a_4 as referrals. As a_1 and a_4 has been contacted before, a_s does not contact these two referrals.

Truster a_s contacts both referrals a_7 and a_8 returned by a_4 . Agent a_7 returns a_1 and a_8 as referrals, but both of them have been contacted before. Truster a_s further contacts the two referrals returned by a_8 , i.e. a_9 and a_{10} . Agent a_{10} is a witness, while agent a_9 is not. Agent a_9 returns a_{11} and a_{12} as referrals. However, since the referral chain already reaches the threshold

(i.e. $U = 4$ in this example), a_s will not contact them.

Hence, in this example, the truster agent a_s manages to find two relevant witnesses, i.e. agent a_6 and a_{10} . The corresponding referral chains are $a_2 \rightarrow a_6$ and $a_1 \rightarrow a_4 \rightarrow a_8 \rightarrow a_{10}$. ■

It should be noted that in the referral process we implicitly assume that agents contacted by the truster a_s are always willing to help a_s find the relevant witnesses. That is, they will return some referrals if they are not the witnesses of the trustee. In MAS where the participating agents usually represent different stakeholders with self-interests, this assumption will not always hold. However, in this thesis, we do not consider how it is guaranteed since it is beyond the control of the trust model and credibility model and highly depends on the application domain considered.

Compared with the original referral process [YS02, YSS04], the credibility-aware referral process brings the following advantages:

- Instead of contacting every referral, the truster only contacts a subset of the referrals. This helps to further reduce the communication cost.
- As the truster agent has the full control over the referral process, it can adaptively direct the referral process. Agents with higher credibility have a higher probability to be contacted. This helps the truster to find more credible witnesses, which also helps to reduce the adverse effects of unfair testimonies.
- Truster agent will also explore the witness agents with lower credibility, though with a lower probability. This gives those less credible agents an opportunity to promote their credibility in the eye of the truster. It also serves as load-balancing, which protects the more credible agents from receiving too many testimony requests.

6.3 Revamping the Referral Process with Modeling of Referral-Credibility

Despite the merits of the credibility-aware referral process, its dependence on the referrers' reports of the referrals' credibility induces the presence of *malicious referrers*. That is, there are

agents who do not give unfair testimonies themselves, but refer the trusters to those referrals that may give unfair testimonies by giving unfairly higher reports of those referrals' credibility. The presence of malicious referrers would intensify the adverse effect of unfair testimonies, since the malicious referrers' credibility reports would direct the "flow" of the trust information (e.g. testimonies) to a "polluted" source. Therefore, to make good use of the credibility-aware referral process, the presence of a malicious referrer must be addressed.

Our strategy to address the presence of malicious referrers is to measure the referral-credibility of each referrer to capture its reliability in reporting the referrals' credibility, and to discount the referrer's reported credibility of the referral with that referrer's referral-credibility. With the referrer's referral-credibility taken into account, instead of contacting a referral agent a_t with a probability of $\Pr(v_{s,t})$, the truster agent a_s contacts the referral agent a_t returned by referrer agent a_r with the following probability

$$MT_{s,r} \cdot \Pr(v_{s,t}) \quad (6.2)$$

Here, $\Pr(v_{s,t})$ is derived with Eq. (6.1), and $MT_{s,r}$ is truster agent a_s 's evaluation of agent a_r 's referral-credibility as a referrer. $MT_{s,r}$ is applied to reduce the probability that the referrals returned by the malicious referrers are visited, and is defined in the range of $[0, 1]$.

An example referral process to find relevant witnesses using referral-credibility is shown in Figure 6.4 (only a snippet of the referral chains is shown in the figure).

In Figure 6.4, a_s is the truster agent who is finding relevant witnesses on the trustee (not shown in the figure). Each edge pointing from a referrer to its referral is labeled with the probability that the referral would be contacted by the truster agent. It is noted that the edges connecting the truster agent a_s with its neighbors are not labeled. It is because the truster agent will contact its neighbors first to find relevant witnesses.

6.4 Modeling Agent's Referral-Credibility

Intuitively, the referral-credibility of an agent can be modeled with the proposed credibility model presented in Chapter 4. Similar to how the proposed credibility model was applied in the context

Figure 6.4: Referral process with the incorporation of referrer's *referral-credibility*

of ART testbed (see Section 5.8.3), the referral-credibility can also be considered the truster's subjective evaluation of the agent's reports on the referrals' credibility. In the proposed credibility model, the credibility of an agent is updated if (1) this agent has given a testimony for the truster's pre-interaction evaluation of overall trust, and (2) the truster updates its post-interaction evaluation of direct trust on the trustee with the new observation of the trustee's behavior in the interaction. To make the credibility model applicable here, two similar conditions must be satisfied. That is, (1) the referrer returns a referral (including this referral's credibility) to the truster, and (2) the referral's credibility can be directly evaluated after the truster contacts the referral. However, a witness is generally found by following more than one referral along a referral chain. Thus it cannot be guaranteed that a referral would always be the witness of the trustee. Hence, condition (2) will not always hold. Given this, the credibility model presented in Chapter 4 is not applicable to model the referral-credibility.

Moreover, because the truster agent finds a witness by following a certain referral chain, which contains a series of referrals, the referral process would be manipulated if any agent along the referral chain gives unfair reports of its referrals' credibility. Therefore, the referral-credibility

of all the referrer agents along the referral chain should be updated each time when the truster agent obtains a new observation regarding the witness agent's credibility. In other words, the update of the referral-credibility must be propagated to all the referrer agents along the referral chain. Such a propagation can naturally be modeled as a *spreading activation* process.

Spreading activation was originally proposed in cognitive science [CL75], and has since then been adapted in other fields, e.g. [ZL04, Cre97]. Generally speaking, it is about determining the *energy* of each node within a network of a number of interconnecting nodes. After injecting an initial *activation* I^0 into a specific node, the node will fan out part of the activation to the nodes it is connected to. Those nodes then iteratively fan out the activation to the other nodes they are connected to. In doing so, the activation spreads over the whole network. If we use $ei(m)$ to denote the activation that a node m receives, the node retains $(1 - d) \cdot ei(m)$ to update its current energy, and fans out the rest $d \cdot ei(m)$ to the nodes it directly connects. The activation that a node n receives from node m is determined as:

$$ei(n) = d \cdot ei(m) \cdot \frac{E(m, n)}{\sum_{(m, s) \in \mathbf{OE}} E(m, s)} \quad (6.3)$$

where d is the *spreading factor*, $ei(m)$ is the activation that node m receives, $E(m, n)$ is the weight of the edge from node m to n , and $\mathbf{OE}$ is the set of node m 's outgoing edges.

When the spreading activation process is applied to model the referral-credibility, the counterpart of the node is each individual referrer agent along the referral chain, and the energy that each referrer agent has is essentially its referral-credibility in the eye of the truster. The activation each agent receives during the course of propagation is in fact the magnitude of the update of the referral-credibility that each agent should undertake. Each agent's referral-credibility is initialized as 1.

The update of the referral-credibility is carried out on the *reversed* referral chain. This is because the update of the referral-credibility is triggered each time the truster agent updates the credibility of the witness agent, which is at the end of the referral chain. Therefore, the initial activation of the spreading activation process is fed in from the end of the referral chain, and propagates backward along the referral chain.

Figure 6.5: Examples of the reversed referral chains

The reversed referral chain is constructed by retaining the edges in the original referral chain and then reversing the directions of the edges. Figure 6.5 shows examples of reversed referral chains based on the referral chains in Figure 6.4. For example, there are outgoing edges from agent a_2 to agent a_5 and from agent a_5 to agent a_7 in Figure 6.4. The corresponding reversed referral chain contains an outgoing edge from agent a_7 to a_5 and another outgoing edge from agent a_5 to agent a_2 .

The quantity of the initial activation I^0 is the absolute change in witness agent a_w 's credibility after truster agent a_s obtains its post-interaction direct trust on trustee agent, i.e. $I^0 = |TT'_{s,w} - TT_{s,w}|$, where $TT'_{s,w}$ and $TT_{s,w}$ are the credibility of the witness agent a_w after and before truster agent a_s obtains its post-interaction direct trust on trustee agent respectively. And the initial activation I^0 is fed into the immediate successor of the witness along the reversed referral chain (i.e. the direct referrer of the witness along the original referral chain). We use $ei(r)$ to denote the activation that a referrer agent a_r receives. Truster agent a_s retains $(1 - d)$ of the activation to update referrer agent a_r 's current referral-credibility, and spreads the rest $d \cdot ei(r)$ along the

reverse referral chain. By retaining $(1 - d) \cdot ei(r)$ to update its referral-credibility, agent a_r 's referral-credibility in the eye of the truster a_s ($MT_{s,r}$) is updated as follows:

$$MT'_{s,r} = (1 - (1 - d) \cdot ei(r)) \cdot MT_{s,r} \quad (6.4)$$

Since each reversed referral chain contains no branches, the third term in the right-hand-side (RHS) of Eq. (6.3) is always 1, i.e. $\frac{E(m,n)}{\sum_{(m,s) \in \mathbf{OE}} E(m,s)} = 1$. Hence the activation that spreads from agent a_r flows completely into its immediate successor along the reversed referral chain. For agent a_r 's immediate successor, the truster agent a_s will repeatedly retain $(1 - d)$ of the activation to update its referral-credibility, and then spread out the rest of the activation. The propagation stops at the last referrer in the reversed referral chain. For this referrer a_r , truster agent a_s will retain all the activation to update its referral-credibility:

$$MT'_{s,r} = (1 - ei(r)) \cdot MT_{s,r} \quad (6.5)$$

The procedure to update the referral-credibility of all the referrer agents along a referral chain is summarized in Figure 6.6.

An example of referral-credibility update is shown below:

Example 6.2 Here, we re-visit Example 6.1. In the referral process shown in Figure 6.3, truster agent a_s finds two relevant witnesses, i.e. agent a_6 and agent a_{10} . The corresponding referral chains are $a_2 \rightarrow a_6$ and $a_1 \rightarrow a_4 \rightarrow a_8 \rightarrow a_{10}$. The update of referrers' credibility is carried out on two reversed referral chains shown in Figure 6.7. The other agents that are not involved in the referral-credibility update are not shown in the figure.

For the reversed referral chain $a_6 \rightarrow a_2$, the initial activation is fed into agent a_2 since a_2 is the direct referrer of the witness agent a_6 . The initial activation in this reversed referral chain is $I_1^0 = |TT'_{s,6} - TT_{s,6}|$. Since agent a_2 is the only referrer along this referral chain, a_s retains all the activation to update its referral-credibility. Hence, its referral-credibility is updated as $MT_{s,2} = (1 - I_1^0) \cdot MT_{s,2}$.

For the reversed referral chain $a_{10} \rightarrow a_8 \rightarrow a_4 \rightarrow a_1$, the initial activation is fed into agent a_8 , and the quantity of the initial activation is $I_2^0 = |TT'_{s,10} - TT_{s,10}|$. Agent a_s retains $(1 - d)$ of

Procedure MT_Update(s, w)

a_s : the truster agent,

a_o : the trustee agent,

a_w : a witness of trustee agent a_o ,

$TT_{s,w}$: agent a_s 's original evaluation of witness a_w 's credibility,

$TT'_{s,w}$: agent a_s 's new evaluation of witness a_w 's credibility after a_s obtains its post-interaction direct trust on trustee agent a_o ,

$\mathcal{R}$: the series of referrers along the referral chain that leads truster agent a_s to witness agent a_w , and they are sorted in the order of appearance in the corresponding *reversed* referral chain,

$|\mathcal{R}|$: the number of referrers in $\mathcal{R}$,

a_r : a referrer in $\mathcal{R}$, r is its index in the sorted series of $\mathcal{R}$,

$MT_{s,r}$: agent a_s 's evaluation of agent a_r 's referral-credibility,

d : the spreading factor.

1: $I^0 = |TT'_{s,w} - TT_{s,w}|$

2: $ei(0) = I^0$

3: **for** $r = 1$ to $|\mathcal{R}|$ **do**

4: $ei(r) = ei(r - 1)$

5: **if** $r == |\mathcal{R}|$ **then**

6: $MT_{s,r} = (1 - ei(r)) \cdot MT_{s,r}$ {this is the last referrer along the reversed referral chain}

7: **else**

8: $MT_{s,r} = (1 - (1 - d) \cdot ei(r)) \cdot MT_{s,r}$

9: $ei(r) = d \cdot ei(r)$

10: **end if**

11: **end for**

Figure 6.6: Spreading activation based referral-credibility update

the activation to update the credibility of agent a_8 and agent a_4 , and spreads out the rest to the next agent along the reversed referral chain. Each edge connecting the two agents is labeled with the activation propagating between the two agents. referral-credibility of agent a_8 and agent a_4 are updated respectively as

$$MT_{s,8} = (1 - (1 - d) \cdot I_2^0) \cdot MT_{s,8};$$

$$MT_{s,4} = (1 - (1 - d) \cdot d \cdot I_2^0) \cdot MT_{s,4}$$

Agent a_1 is the last referrer in this reversed referral chain. Hence, agent a_s retains all the activation to update its referral-credibility: $MT_{s,1} = (1 - d \cdot d \cdot I_2^0) \cdot MT_{s,1}$. ■

Figure 6.7: Example of referral-credibility update

6.5 Empirical Evaluation

Having presented the modeling of agent's referral-credibility as well as the credibility-aware referral process, we now turn to the evaluation of the modeling of referral-credibility.

6.5.1 Experimental Setup and Methodology

A number of experiments has been conducted to evaluate the effectiveness of the referral-credibility model. The experiments are again conducted in `TrustBed`. The default settings of the testbed parameters are applied as shown in Table 5.4. The credibility model used by the consumer agents retains 10 as the Z value.

The referral process to find the relevant witnesses is controlled by the following parameters:

- The number of neighbors that each agent maintains, i.e. N_h .
- The threshold of the referral chain length, i.e. U .

Generally speaking, N_h limits the breadth of the range in which the consumer agents will search for relevant witnesses, and U limits how far the agents will go for the relevant witnesses. These two parameters in combination define the search range of the referral process. Larger values of N_h and U result in a broader and deeper search range for each individual consumer

agent. This generally leads to a better performance than smaller one since more testimonies are made available as evidence for the agent to make the trust evaluations. Nevertheless, as a trade-off, a broader and deeper search also requires more resources (e.g. both in terms of memory, waiting time, communication costs and etc.) from the consumer agents. In light of this, these two parameters are usually set with moderate values. In the experiments, it is set that $N_h = 3$ and $U = 6$. The referral process has already been applied in the empirical evaluations in the Chapter 5, and same settings of N_h and U are applied.

In the experiments, the malicious referrers are implemented to present together with the presence of collusive unfair testimonies. This is due to two considerations. First, collusive unfair testimonies are more successful in cheating the proposed credibility models. Second, with the presence of collusion, the malicious referrer has a clear target whose credibility will be manipulated in its referrals reports. That is, it will give unfairly higher reports of the credibility for those colluding consumers to make them more likely to be contacted by the trusters. When a referrer a_r generates a referral, it will reveal its evaluation of the referral's credibility without manipulation if it is not a malicious referrer. Otherwise, it will give an unfairly higher report of the referral's credibility if the referral is a colluding consumer, which is implemented as $0.5 + \max(TT_{r,t})$, where $\max(TT_{r,t})$ denotes the maximum of the credibility of all the referrals returned by referrer a_r . The report of the referral credibility after manipulation is clamped in the range of $[0, 1]$.

With the presence of malicious referrers, the collusion with Bad providers (see Table 5.2 for the definition of Bad provider) contains two types of consumers. One type of consumers are witnesses who give **ballot-stuffing** unfair testimonies on the colluding providers, while another type of consumers are malicious referrers who give unfairly higher reports on the colluding witnesses' credibility in order to make them more possible to be contacted by the trusters.

We study the effectiveness of the referral-credibility model in the presence of various combinations of collusive unfair testimonies and malicious referrers. A methodology similar to the one used in Section 5.3 is applied here. That is, the consumers are separated into two groups. One group of consumer agents are equipped with the referral-credibility model (denoted as group Meta), while the other group of consumers are not (denoted as group NoMeta). Each group is

again populated with $N_c = 100$ consumer agents, and both groups are applying the same basic trust model and the proposed credibility model. Then the effectiveness of the referral-credibility model is investigated by comparing the performance of the consumer agents in the two groups. The performance of different groups is measured in terms of *collusion power*, which is defined in Section 5.3.2. This metric basically evaluates how many times the non-colluding consumers in each group have tried the colluding providers' services to solve their problem. It is noted that the collusion power is only calculated on the non-colluding consumers. For instance, if the configuration of collusive unfair testimonies is Pos60, and 20% of all the consumers are malicious referrers, the collusion power is evaluated on the other 20% of the consumer agents.

6.5.2 The Results

Figure 6.8 presents the performance of group Meta and NoMeta in the presence of various combinations of malicious referrers and collusive unfair testimonies. The first thing this figure shows is that the presence of malicious referrers has increased the *collusion power*, as compared with the results obtained in previous chapter (see Figure 5.4 (b)). In Figure 5.4 (b), group Cred (without referral-credibility model, i.e. group NoMeta here) is able to achieve a *collusion power* of less than 0.25 (i.e. try the colluding providers' services once for every 4 problems encountered) in the presence of various configurations of collusive unfair testimonies but without malicious referrers. However, Figure 6.8 shows that the presence of 20%, 40% and 60% malicious referrers have increased the collusion power of group NoMeta to higher than 0.3, 0.35, and 0.5 respectively. This observation confirms that the presence of malicious referrers does exacerbate the adverse effect of unfair testimonies. This is because the malicious referrers increase the probability that a truster contacts a colluding consumer who gives unfair testimonies on Bad providers. Consequently, the unfair testimonies will have a higher presence in the testimonies that are collected by the trusters. Since the collusive unfair testimonies are more successful in cheating the credibility model, a higher presence of collusive unfair testimonies will adversely manipulate the non-colluding consumers' evaluations of trust on the colluding Bad providers, which results in a higher collusion power.

(a)

(b)

(c)

Figure 6.8: Performance of different groups in the presence of various combination of malicious referrers and collusive unfair testimonies: (a) 20% malicious referrers; (b) 40% malicious referrers; (c) 60% malicious referrers;

On the other hand, consumers in group *Meta*, equipped with the referral-credibility model, are able to mitigate the influence of malicious referrers. With the modeling of referral-credibility, consumer agent's referral-credibility decreases every time it is found to give a unfair report of its referrals' credibility. By discounting the referrers' future reports of referrals' credibility by the referrers' referral-credibility, the probability that the colluding consumers would be contacted by the trusters is reduced. This consequently accounts for the reduction of the presence of unfair testimonies in the testimonies collected by the trusters.

It is noted that the referral-credibility model only reduces the probability that the colluding consumers would be contacted during the referral process. The colluding consumers would still be contacted by the non-colluding consumers, though with a lower probability. Hence, unfair testimonies still exert influence on the non-colluding consumers' evaluations of trust on the colluding providers. Nevertheless, results in Figure 6.8 show that the deployment of referral-credibility model is still able to maintain group *Meta*'s performance (in terms of collusion power) at a comparable level to that in the case without presence of malicious referrers.

The referral-credibility update process has one parameter, i.e. the spreading factor d in Eq. (6.4). With a smaller value of d (i.e. $d < 0.5$), the truster agent tends to assign more penalty to referrers close to the witnesses, while with a larger of d (i.e. $d > 0.5$), the truster agent tends to give more penalty to referrers far away from the witnesses. The above results are obtained with $d = 0.5$. That is, truster agent retains 50% of the activation that each referrer receives to update the corresponding referrer's referral-credibility. $d = 0.5$ shows a balance between the aforementioned two cases. Other values of d (both $d < 0.5$ and $d > 0.5$) have also been applied in the experiments. Similar trends are observed, which shows the effectiveness of the referral-credibility modeling in various configurations.

6.6 Summary

One of the fundamental issues in the research on trust modeling in MAS is to mitigate the adverse effect of unfair testimonies [JIB07]. Previous chapters (Chapter 3, 4, and 5) have presented our proposal to address this issue. An implicit assumption of previous chapters is that the truster

The Complete Trust Model

a_s : the truster agent,

a_o : the trustee agent,

a_w : a witness of the trustee agent a_o ,

$TT_{s,w}$: agent a_s 's evaluation of witness a_w 's credibility,

$TP_{s,o}$: agent a_s 's pre-interaction direct trust on agent a_o ,

$TA_{s,o}$: agent a_s 's (pre-interaction) evaluation of overall trust that it should place in trustee a_o .

$TP'_{s,o}$: agent a_s 's post-interaction direct trust on agent a_o with new observation of a_o 's behavior in the interaction.

- 1: a_s builds and maintains each witness agent a profile in giving testimonies
 - 2: a_s evaluates $TP_{s,o}$ with past personal interaction experience {any trust model could be employed to make the trust evaluations here}
 - 3: Agent a_s finds relevant witnesses with the credibility-aware referral process (see Figure 6.2)
 - 4: Let $\mathcal{W}_{s,o}$ = the set of witnesses on the trustee agent a_o located in the previous step
 - 5: **for all** witness agent $a_w \in \mathcal{W}_{s,o}$ **do**
 - 6: a_s evaluates witness agent a_w 's credibility $TT_{s,w}$ using Eq. (4.2)
 - 7: a_s adjusts the testimony given by a_w according to Eq. (4.4) if its testimony is not filtered
 - 8: **end for**
 - 9: a_s aggregates all the adjusted testimonies and $TP_{s,o}$ using Eq. (4.6) to evaluate $TA_{s,o}$
 - 10: **if** a_s interacts with trustee agent a_o after the pre-interaction trust evaluation **then**
 - 11: a_s evaluates $TP'_{s,o}$
 - 12: **for all** witness agent $a_w \in \mathcal{W}_{s,o}$ **do**
 - 13: **if** a_w 's testimony has been aggregated by a_s in the pre-interaction trust evaluation **then**
 - 14: a_s updates a_w 's credibility $TT_{s,w}$ with $TT_Update(s, w)$ as shown in Figure 4.5
 - 15: a_s updates the referral-credibility of the referrers along the referral chain leading to witness a_w using $MT_Update(s, w)$ as shown in Figure 6.6
 - 16: **end if**
 - 17: **end for**
 - 18: **end if**
-

Figure 6.9: The Complete Trust Model

agent would ask for testimonies from all other agent in MAS. This assumption may not always hold. This chapter further improves our proposal by discarding this assumption.

This chapter proposes a credibility-aware referral process to help agents to find relevant witnesses in an efficient manner. With the credibility-aware referral process, each agent maintains direct connection with a number of agents, and asks those agents to find relevant witnesses for it iteratively. Moreover, the truster agents are able to adaptively direct the witness discovery process toward those credible agents. On the other hand, this opens the possibility of presence of malicious referrers, who do not give unfair testimonies themselves but direct the truster agent to other agents who do give unfair testimonies. The presence of malicious referrer aggravates the

adverse effect of unfair testimonies. To address the presence of malicious referrers, we model the referral-credibility of each agent to capture its reliability in reporting other agents' credibility. The empirical studies validate the effectiveness of the modeling of agent's referral-credibility in combating the malicious referrers.

The proposed credibility model and referral-credibility model are generic and can be applied to most computational trust models. With incorporation of these two models to combat unfair testimonies and malicious referrers, a complete trust model operates as shown in Figure 6.9.

This chapter has concluded the research of making the trust models robust in the presence of unfair testimonies. The next chapter summarizes the contributions of this thesis and outlines the directions for future work.

Chapter 7

Conclusions

Trust is an ubiquitous concept that underlies each interaction in human society. Due to its importance, trust has long been a research topic in various disciplines such as sociology, economics, and philosophy. In recent years, a large body of research work has emerged on trust modeling in MAS [BK02, RHJ04, JIB07]. Generally, trust models help one agent to evaluate its trust on another agent primarily with two sources of information:

- (a) an agent's personal interactions experience with another agent;
- (b) third-party testimonies about their past interactions with another agent.

The problem arises when third-party testimonies are taken into account. That is, there exists the presence of *unfair testimonies*. The presence of unfair testimonies would deteriorate the effectiveness of the computational trust models [JIB07, Del00, WJI04].

Mitigating or even avoiding the adverse effect of unfair testimonies is a problem that every computational trust model should address [JIB07]. This chapter summarizes our efforts in addressing this problem. Section 7.1 reviews the contributions of this thesis. Following that, Section 7.3 discusses the possible directions in which this research can be continued in future.

7.1 Research Contributions

This thesis devotes itself to making the trust model robust in the presence of unfair testimonies. By accomplishing this goal, this thesis has made the following contributions:

- (i) As a starting point, Chapter 3 proposes a testimony-filtering method. The filtering method identifies the possible unfair testimonies by measuring the uncertainty that each testimony contains. Then the remaining testimonies are aggregated with the truster's own evaluations of direct trust on the trustee.

Compared with the existing work, the proposed filtering method does not require any additional mechanism or knowledge other than the testimonies exchanged between the agents. Moreover, the proposed uncertainty-based filtering method outperforms other related work, e.g. a typical *endogenous* filtering method proposed by Whitby et al. [WJI04] both in terms of processing time and effectiveness to mitigate the adverse effect of unfair testimonies.

- (ii) The uncertainty-based filtering method is then revamped in Chapter 4 with a credibility model. With the proposed credibility model, the truster measures the credibility of each agent, which essentially evaluates the usefulness of this agent's testimonies.

The proposed credibility model preserves the merits of the uncertainty-based filtering method. For example, it does not require any additional mechanism or knowledge other than the testimonies exchanged among agents. Apart from preserving the filtering method's merits, the credibility model improves the filtering method in many aspects and has its own merits. It improves the uncertainty-based filtering method with a more adaptive filtering criterion. It also improves the filtering method with a method to adjust the testimony which makes a testimony useful even it is a possible unfair testimony. The empirical evaluations in Chapter 5 also show that it consistently outperforms related work in all the experimental settings considered.

Another significant merit of the credibility model is that it is decoupled from the trust model. It does not explicitly require a particular trust model to be used. Although the `Beta Reputation System` is employed as the basic trust model in this thesis, other computational trust models can also be used as long as the outcomes of the trust evaluations can be represented in the form of discrete ratings. This makes it a generic solution to tackle the presence of unfair testimonies.

- (iii) Unlike the existing work, which assumes that agents would ask all the other agents in MAS for testimonies, this thesis makes no such assumption. Chapter 6 proposes a credibility-aware referral process to facilitate agents to find relevant witnesses efficiently. During the referral process, there are *malicious referrers* who direct agents to others who are giving unfair testimonies. The presence of malicious referrers aggravates the adverse effect of unfair testimonies. However, the presence of malicious referrers are generally left undressed in current research on trust modeling in MAS due to the naive assumption that the agents would contact directly all the other agents when discovering testimonies. In Chapter 6, an approach based on the theory of *spreading activation* is proposed to counteract the presence of malicious referrers. To the best of our knowledge, it is the first to address this issue in research on trust modeling in MAS.

7.2 Potential Applications

It is reiterated that the proposed research can be applied to many application domains. Some typical application domains include Peer-to-Peer content-sharing networks, Grid, and massive multi-player online role playing games (MMORPGs).

7.2.1 Application in P2P Network and Grid

The real problem underlies the Peer-to-Peer content-sharing networks and Grid is resource sharing among various users [FKT01]. Generally, the reliability of the providers who manage the resources (either digital files or other computing resources) are unknown to the users before utilizing the resources, i.e. the users who manage the resources may fail to provide the resources with the quality they agreed to deliver. For example, in the context of service-oriented Grid, a service provider may intentionally discard a user's computing request in favor of another user's request as the latter can pay more. It is desired that users have the ability of making prudent resource selection decision regarding the reliability before utilizing the resource. Establishing trust within P2P networks and Grid is an alternative to meet this requirement. There is also a

number of existing attempts to achieve this goal, e.g. [DdVP⁺02, AVvLA03, KSGM03]. However, the existing work is usually built on an assumption that all the witnesses would not share unfair testimonies. There is a lack of mechanism to address the unfair testimonies, which renders the existing work susceptible to the presence of unfair testimonies. The research reported in this thesis can be applied to enhance the existing attempts' robustness in the presence of unfair testimonies.

When this research is applied in the context of P2P network and Grid, each user within the system is modeled as an agent. Each agent is allowed to give feedbacks for its interaction partner. Based on these feedbacks, the trustworthiness of the interaction partner can be evaluated with any of the existing attempts, e.g. [DdVP⁺02, AVvLA03, KSGM03]. Subsequently, the proposed credibility model can be incorporated to evaluate the credibility of other testimony sources. As discussed in Chapter 4, the credibility model also filters the third-party testimonies and aggregates the third-party testimonies, which contributes to mitigate the adverse influence of unfair testimonies and make the trust evaluations more accurate. Moreover, thanks to the distributed nature of P2P network and Grid, the modeling of referral-credibility can easily be incorporated to further reduce the adverse effect of unfair testimonies.

7.2.2 Application in MMORPG

MMOPRG is often afflicted with problems such as dishonesty, theft, and malice [Smi04]. Most of the current MMORPGs provide the mechanisms for the players to manually block or ignore certain players if the former consider the latter are harassing them. However, such mechanisms are not very useful when players are stuck halfway in a raid or in completing a quest. It is preferred that the players are able to automatically block the other players from contacting them based on the latter's past behaviors, such as the trust rating system built into the Xbox Live [xbo07]. Establishing trust within the game is a promising approach to meet this requirement [Smi04].

When this research is applied in the context of MMORPG, each player avatar is basically an agent, who is responsible for evaluating the trustworthiness of other player avatars. Furthermore, with the proposed credibility and referral-credibility model applied, each player avatar also

measures the credibility and referral-credibility of other player avatars. This helps to improve the accuracy of trust evaluations, which consequently leads to a better identification of “bad” players.

7.3 Future Directions

There are several areas in which current research can be extended:

- (i) Agent may also change its behavior in giving testimonies over time. The current design of the credibility model is able to capture such behavior change during a witness agent’s existence in MAS by storing all its past testimonies in its profile. Furthermore, the model can evaluate credibility based on the full history of the witness’ past behaviors. However, in some application domains, agents may not be so prudent in evaluating the witnesses. For those agents, a witness that is credible within a relatively shorter period of time would suffice. In this case, when evaluating a witness agent’s credibility, the credibility model should maintain a “shorter memory” of a witness’ past testimonies. However, the current design of the credibility model records the witness’ behavior in giving testimony by updating the counting information in the witness’ profile. There is no temporal information recorded in the witness’ profile, namely, it does not record when the witness has given a testimony. Therefore, it is not able to forget a particular testimony given by the witness in the past by forgetting a past testimony as in the basic trust model presented in Section 3.2 (see Eq. (3.4)).

One possible way to equip the credibility model with the ability to gradually forget a witness’ past behavior would be as follows: besides maintaining the counting information, each cell in the profile also needs to store the time of each corresponding testimonies. Then before each update to the profile, old testimonies beyond a predefined time window will be discarded. Then the credibility is calculated on the updated profile. In order to enable such modification, the current design of the profile needs to be significantly extended to incorporate the temporal information. And such extension would definitely induce extra storage overhead for the agents. However, such an extension is expected to enhance the

credibility model's applicability in above-mentioned domains where "shorter memory" of the witnesses is desired.

- (ii) The next potential extension is with the counteraction of the presence of malicious referrers. In the current research, the influence of the malicious referrers is addressed by modeling the agents' referral-credibility. Incorporation of referral-credibility into the credibility-aware referral process reduces the possibilities that colluding witnesses could be contacted by the trusters, which accounts for the reduction of the presence of unfair testimonies in all the available testimonies collected by the trusters. Although the results of the empirical evaluations validate the effectiveness of the modeling of agent's referral-credibility, there is still room for improvement. The counteraction of the presence of malicious referrers can be further improved by building a (weighted) graph to represent the relationships between all the agents involved (including the referrers, referrals, and the trustees) during the referral processes, and then applying *social network analysis* [WF94] to discover the hidden collusion among the agents. With the hidden collusion detected, the trusters can then focus only on reducing the probability that the colluding witnesses would be contacted. This is preferable to reducing the probability that all the referrers returned by a malicious referrer would be contacted. The result of the collusion detection can in turn contribute to reducing the adverse effect of unfair testimonies.
- (iii) All the work discussed above focuses on extending the capability of the credibility model or referral-credibility model. Nevertheless, when the credibility model and mete-credibility model, as well as the trust model, are deployed in real environments, it is not realistic for an agent, say a_s , to select its interaction partner solely based on the evaluations of trust on the candidate partners. Agent a_s may also need to consider other factors like the importance of the interaction, the cost, and the gain. For example, suppose agent a_s is looking for an interaction partner for a not-so-critical task. Assume there are two candidates: one is more trustworthy but asks for higher payment, the other is marginally less trustworthy but asks for lower payment. In this case, the latter might be a better choice. On the other hand, in

another example, with the same two candidates, the task is now very critical to agent a_s . In this case, a_s might select the former since it cannot accept a catastrophic outcome if the selected partner fails to accomplish the task. As these examples show, simply comparing the trustworthiness of candidate partners might not always suffice. Therefore, a decision-making model that takes into account the partners' trustworthiness, the importance and risk of the interaction, the cost, the gain and so on, is obviously desirable.

- (iv) Another promising future work is the design of a special type of agent, namely *trust agent*. By using the term *trust agent*, it means that: other than the basic characteristics that an agent has, e.g. ability to adapt and ability to learn, agent also has (1) the ability to evaluate potential interaction partners' trustworthiness, and (2) the ability to make trust-aware decision regarding the selection of interaction partner before engaging in an interaction.

Trust agent acquires the first ability with the application of trust models, which evaluate trustworthiness based on the interaction partner's past behavior. The research presented in this thesis, as well as the future direction (i) and (ii) shown above, contribute to enhance the performance of the trust models in the presence of unfair third-party testimonies. Trust agent's second ability is enabled with the decision-making model mentioned above in (iii).

It should be noted that *trust agent* makes the trust-aware interaction partner selection before any interaction. Although this would incur overhead, both in terms of computation (for evaluating the trustworthiness) and communication (for collecting the testimonies), the use of *trust agent* appears to offer certain advantages. An obvious advantage is that it increases agent's confidence in interacting with others, reduces agent's utility loss, and enhances agent's applicability in MAS, in which each agent maximizes its own gain even at the cost of others.

Appendix A

Credibility-modeling Method proposed by Yu and Singh

The method proposed by Yu and Singh [YSS04, YS02] restricts agent's credibility to be in the range of $[0, 1]$, and initializes each individual agent's credibility to be maximum value, i.e. 1. Then the credibility is updated with a variant of the Weighted Majority Algorithm [LW94]. If the last testimony given by witness a_w is similar to truster a_s 's direct trust in trustee a_o after interacting with that trustee, witness a_w 's credibility remains unchanged. Otherwise, as a penalty, a_w 's credibility is decreased with the following formula:

$$TT'_{s,w} = TT_{s,w} \cdot \theta \quad (\text{A.1})$$

θ denotes the amount of the penalty. It is determined according to the distance between a_w 's testimony and a_s 's post-interaction direct trust in the same trustee, and it can be any value that satisfies:

$$\delta^{|TP_{w,o} - TP_{s,o}|} \leq \theta \leq 1 - (1 - \delta)|TP_{w,o} - TP_{s,o}| \quad (\text{A.2})$$

$TP_{w,o}$ and $TP_{s,o}$ denote witness a_w 's testimony and truster a_s 's post-interaction direct trust on the trustee a_o . δ is a constant. In the empirical studies in this thesis, θ is chosen as the upper bound of the range, i.e. $\theta = 1 - (1 - \delta)|TP_{w,o} - TP_{s,o}|$, and δ is set as 0.5. This setting is consistent with the empirical value as reported in [YSS04].

Then the testimonies are aggregated with the truster a_s 's evaluation of direct trust in the following manner:

Suppose $\{w_1, w_2, \dots, w_L\}$ are a group of witnesses on trustee a_o , then the aggregation of testimonies is done as:

$$\mathcal{P} = \begin{cases} \sum_{k=1}^L TT_{s,w_k} * TP_{w_k,o} / L & L \neq 0 \\ 0.5 & L = 0 \end{cases} \quad (\text{A.3})$$

The overall trust that a_s should place in trustee a_o is:

$$TA_{s,o} = \begin{cases} \eta * TP_{s,o} + (1 - \eta) * \mathcal{P} & L \neq 0 \\ 0.5 & L = 0 \end{cases} \quad (\text{A.4})$$

where L is the number of the witnesses that have reported testimonies, η is truster a_s 's confidence about its evaluation of direct trust in trustee a_o . It is calculated as:

$$\eta = h/H \quad (\text{A.5})$$

where H is the predefined number of ratings that are taken into account when a_s evaluating the direct trust in a_o (like the parameter of W in Eq. (3.4)), and h is the number of a_s 's past interactions with a_o . Eq. (A.4) implies that a_s 's evaluation of direct trust in trustee a_o becomes more and more important in evaluating the overall trust in a_o , and when $h \geq H$, third-party testimonies are not aggregated at all.

Appendix B

List of Publications

The work reported in this thesis has contributed a total of nine international publications, including a few other awaiting review results. The list is given below.

Journal Papers

- (1) **Jianshu Weng**, Zhiqi Shen, Chunyan Miao, and Angela Goh, A General Credibility Model to Cope with Unfair Testimonies. Submitted to *IEEE Transactions on Knowledge and Data Engineering*, 2007.
- (2) **Jianshu Weng**, Chunyan Miao, and Angela Goh, An Entropy-based Approach to Protecting Rating Systems from Unfair Testimonies. *IEICE Transaction on Information and Systems*, Vol. E89-D, No 9, pp 2502-2511, 2006.
- (3) Chunyan Miao, **Jianshu Weng**, Angela Goh, Zhiqi Shen, and Bo An, Fuzzy cognitive maps for dynamic grid service negotiation. *Multiagent and Grid Systems*, Vol. 2, No 2, pp 101-114, 2006.

Conference Papers

- (1) **Jianshu Weng**, Chunyan Miao, and Angela Goh, A General Credibility Model to Cope with Unfair Testimonies. In *Proceedings of 1st International Conference on Ubiquitous Information Management and Communication*, Suwon, Korea, pp 255-264, 2007. **Invited Paper.**

- (2) **Jianshu Weng**, Chunyan Miao, and Angela Goh, A Robust Reputation System for the Grid. In *Proceedings of the 6th IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2006)*, Singapore, pp 548-551, 2006. (Acceptance rate: 25%)

- (3) **Jianshu Weng**, Chunyan Miao, Angela Goh, Zhiqi Shen, and Robert Gay, Trust-based Agent Community for Collaborative Recommendation. In *Proceedings of the 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006)*, Hakodate, Japan, pp 1260-1262, 2006. (Acceptance rate: 23%)

- (4) **Jianshu Weng**, Chunyan Miao, and Angela Goh, Improving Collaborative Filtering with Trust-based Metrics. In *Proceedings of The 21st Annual ACM Symposium on Applied Computing (SAC 2006)*, Dijon, France, pp 1860-1864, 2006. (Acceptance rate: 32%)

- (5) **Jianshu Weng**, Chunyan Miao, Angela Goh, and Dongtao Li, Trust-based collaborative filtering. In *Proceedings of 2005 ACM International Conference on Information and Knowledge Management (CIKM 2005)*, Bremen, Germany, pp 299-300, 2005. (Acceptance rate: 17.9%)

- (6) **Jianshu Weng**, Chunyan Miao, and Angela Goh, Protecting Online Rating Systems from Unfair Ratings. In *Proceedings of the 2nd International Conference on Trust, Privacy, and Security in Digital Business (TrustBus 2005)*, with DEXA 2005, Copenhagen, Denmark, pp 50-59, 2005.

Bibliography

- [ARH00] Alfarez Abdul-Rahman and Stephen Hailes. Supporting trust in virtual communities. In *Proceedings of 33rd Hawaii International Conference on System Sciences*, volume 1, pages 9–17, 2000.
- [AVvLA03] Beulah Alunkal, Ivana Veljkovic, Gregor von Laszewski, and Kaizar Aminand. Reputation-based grid resource selection. In *Proceedings of Workshop on Adaptive Grid Middleware (AGridM 2003)*, 2003.
- [Bar83] Bernard Barber. *The Logic and Limits of Trust*. Rutgers University Press, 1983.
- [BK95] Ronald S. Burt and Marc Knez. Kinds of third-party effects on trust. *Rationality and Society*, 7(3):255–292, 1995.
- [BK02] K. Suzanne Barber and Joonoo Kim. Soft security: Isolating unreliable agents. In *Proceedings of 5th Workshop on Deception, Fraud, and Trust in Agent Societies, AAMAS 2002*, pages 8–17, 2002.
- [CB90] George Casella and Roger L. Berger. *Statistical Inference*. The Wadsworth & Brooks/Cole Publishing Company, 1990.
- [Ces90] Bojan Cestnik. Estimating probabilities: A crucial task in machine learning. In *Proceedings of 9th European Conference on Artificial Intelligence*, pages 147–149, 1990.
- [CF98] Cristiano Castelfranchi and Rino Falcone. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proceedings of 3rd International Conference on Multi-Agent Systems (ICMAS 1998)*, pages 72–79, 1998.
- [CGJ96] David. A. Cohn, Zoubin. Ghahramani, and Michael. I. Jordan. Active learning with statistical models. *JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH*, 4:129, 1996.
- [CL75] Allan. M. Collins and Elizabeth. F. Loftus. A spreading activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975.

- [Cre97] Fabio. Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–482, 1997.
- [CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 1991.
- [DdVP⁺02] Ernesto Damiani, De Capitani di Vimercati, Stefano Paraboschi, Pierangela Samarati, and Fabio Violante. A reputation-based approach for choosing reliable resources in peer-to-peer networks. In *Proceedings of the 9th ACM Conference on Computer and Communications Security (CCS 2002)*, pages 207–216, 2002.
- [Del00] Chrysanthos Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proceedings of the 2nd ACM Conference on Electronic commerce (EC 2000)*, pages 150–157, 2000.
- [Dem68] Arthur P. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society, Series B*, 30(2):205–247, 1968.
- [Deu62] Morton Deutsch. Cooperation and trust: Some theoretical notes. In Marshall. R. Jones, editor, *Nebraska Symposium on Motivation*, pages 275–319. Nebraska University Press, 1962.
- [FKM⁺05] Karen K. Fullam, Tomas B. Klos, Guillaume Muller, Jordi Sabater, Andreas Schlosser, Zvi Topol, K. Suzanne Barber, Jeffrey S. Rosenschein, Laurent Vercouter, and Marco Voss. A specification of the Agent Reputation and Trust (ART) testbed: experimentation and competition for trust in agent societies. In *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005)*, pages 512–518, 2005.
- [FKM⁺06] Karen K. Fullam, Tomas B. Klos, Guillaume Muller, Jordi Sabater, Andreas Schlosser, Zvi Topol, K. Suzanne Barber, Jeffrey S. Rosenschein, Laurent Vercouter, and Marco Voss. The Agent Reputation and Trust (ART) testbed game description. Available at <http://www.lips.utexas.edu/art-testbed/pdf/SpecSummary.pdf>, 2006.
- [FKT01] Ian Foster, Carl Kesselman, and Steven Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *International Journal of Supercomputer Applications*, 15(3):200–222, 2001.
- [Gam88] Diego Gambetta. Can we trust trust? In *Trust: Making and Breaking Cooperative Relations*, chapter 13. Basil Blackwell, 1988.

- [GK54] Leo A. Goodman and William H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764, 1954.
- [GS00] Tyrone Grandison and Morris Sloman. A survey of trust in Internet application. *IEEE Communications Surveys and Tutorials*, 3(4), 2000.
- [HJL03] Minghua He, Nicholas R. Jennings, and Ho-fung Leung. On agent-mediated electronic commerce. *IEEE Transaction on Knowledge and Data Engineering*, 15(4):985–1003, 2003.
- [HJS06a] Trung Dong Huynh, Nicholas R. Jennings, and Nigel R. Shadbolt. Certified reputation: How an agent can trust a stranger. In *Proceesings of the 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006)*, pages 1217–1224, 2006.
- [HJS06b] Trung Dong Huynh, Nicholas R. Jennings, and Nigel R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Journal of Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.
- [Jen01] Nicholas R. Jennings. An agent-based approach for building complex software systems. *Communication of ACM*, 44(4):35–41, 2001.
- [JF03] Radu Jurca and Boi Faltings. Towards incentive-compatible reputation management. In L. Korba R. Falcone, S. Barber and M. Singh, editors, *Trust, Reputation and Security: Theories and Practice*, volume 2631 of *Lecture Notes in AI*, pages 138–147. Springer-Verlag, 2003.
- [JF04] Radu Jurca and Boi Faltings. Eliciting truthful feedback for binary reputation mechanisms. In *Proceedings of the 3rd IEEE/WIC/ACM International Conference on Web Intelligence (WI2004)*, pages 214–220, 2004.
- [JHF03] Audun Jøsang, Shane Hird, and Eric Faccar. Simulating the effect of reputation systems on e-markets. In *Proceedings of the 1st International Conference on Trust Management (iTrust2003)*, pages 179–194, 2003.
- [JI02] Audun Jøsang and Roslan Ismail. The Beta reputation system. In *Proceedings of the 15th Bled Conference on Electronic Commerce*, 2002.
- [JIB07] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decision Support System*, 43(2):618–644, 2007.

- [JSW98] Nicholas R. Jennings, Katia Sycara, and Michael Wooldridge. A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems*, 1(1):7–38, 1998.
- [KSGM03] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The EigenTrust algorithm for reputation management in P2P networks. In *Proceedings of 12th International Conference on World Wide Web (WWW2003)*, pages 640–651, 2003.
- [LR04] Shyong K. Lam and John Riedl. Shilling recommender systems for fun and profit. In *Proceedings of the 13th international conference on World Wide Web*, pages 393–402, 2004.
- [LW94] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, Feb 1994.
- [Mar94] Stephen Paul Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, Department of Mathematics and Computer Science, University of Stirling, 1994.
- [MC96] D. Harrison McKnight and Norman L. Chervany. The meanings of trust. MISRC Working Paper Series. Available at <http://misrc.umn.edu/wpaper/wp96-04.htm>, 1996.
- [Mit97] Tom Mitchell. *Machine Learning*. McGraw Hill, 1997.
- [Mui03] Lik Mui. *Computational Models of Trust and Reputation: Agents, Evolutionary Games, and Social Networks*. PhD thesis, Department of Electrical Engineering and Computer Science, MIT, 2003.
- [Nat05] National White Collar Crime Center and the Federal Bureau of Investigation. IC3 2004 Internet Fraud-Crime Report. Available at http://www.ifccfbi.gov/strategy/2004_IC3Report.pdf, 2005.
- [OKY06] Özgür Kafali and Pinar Yolum. Trust strategies for ART testbed. In *Proceedings of the Workshop on Trust in Agent Societies at The 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS2006)*, pages 43–49, 2006.
- [PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
- [RHJ04] Sarvapali D. Ramchurn, Dong Huynh, and Nicholas R. Jennings. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19(1):1–25, 2004.

- [RZ02] Paul Resnick and Richard Zeckhauser. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. In Michael R. Baye, editor, *The Economics of the Internet and E-commerce*, volume 11 of *Advances in Applied Microeconomics*, pages 127–157. Elsevier, 2002.
- [SB98] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.
- [Sha76] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.
- [SM03] Jordi Sabater-Mir. *Trust and Reputation for Agent Societies*. PhD thesis, Institut d'Investigació en Intel·ligència Artificial, 2003.
- [Smi04] Jonas Heide Smith. Trusting the avatar. Available at <http://jonassmith.dk/weblog/uploads/TrustingtheAvatardraftsubmitted.pdf>, 2004.
- [Syc98] Katia P. Sycara. Multiagent systems. *AI Magazine*, 19(2):79–92, 1998.
- [Ten05] Jay M. Tenenbaum. AI meets Web 2.0: Building the web of Tomorrow, Today. Technical Report CN-TR-05-07, CommerceNet Labs, 2005.
- [TPJL06] W. T. Luke Teacy, Jigar Patel, Nicholas R. Jennings, and Michael Luck. TRAVOS: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.
- [WF94] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
- [WJ95] Michael Wooldridge and Nicholas R. Jennings. Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10(2):115–152, 1995.
- [WJI04] Andrew Whitby, Audun Jøsang, and Jadwiga Indulska. Filtering out unfair ratings in Bayesian Reputation Systems. In *Proceedings of the Workshop on Trust in Agent Societies at the 3rd International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS2004)*, 2004.
- [Woo97] Michael Wooldridge. Agent-based software engineering. *IEE Proceedings Software Engineering*, 144(1):26–37, 1997.
- [Woo02] Michael Wooldridge. *An Introduction to MultiAgent System*. John Wiley & Sons, 2002.

- [xbo07] Xbox Live player feedback system. Available at <http://www.xbox.com/en-US/support/systemuse/xbox360/livefeatures/feedback.htm>, 2007.
- [YS02] Bin Yu and Munindar P. Singh. An evidential model of distributed reputation management. In *Proceedings of 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS2002)*, pages 294–301, 2002.
- [YS03] Bin Yu and Munindar P. Singh. Detecting deception in reputation management. In *Proceedings of 2nd International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS2003)*, pages 73–80, 2003.
- [YS05] Pinar Yolum and Munindar P. Singh. Engineering self-organizing referral networks for trustworthy service selection. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 35(3):396–407, 2005.
- [YSS04] Bin Yu, Munindar P. Singh, and Katia Sycara. Developing trust in large-scale peer-to-peer systems. In *Proceedings of 1st IEEE Symposium on Multi-Agent Security and Survivability*, pages 1–10, 2004.
- [ZL04] Cai-Nicolas Ziegler and Georg Lausen. Spreading activation models for trust propagation. In *Proceedings of 2004 IEEE International Conference on e-Technology, e-Commerce, and e-Services (EEE 04)*, pages 83–97, 2004.

Index

- agent, 5
- context, 25
- cost, gain, 88, 91
- credibility, 17, 64
- current belief, 49
- decaying factor, 46
- dummy cell, 71
- MAE, *see* mean absolute error
- malicious referrer, 18, 116, 122
- MAS, *see* multi-agent system
- mean absolute error, 53, 78
- multi-agent system, 6
- POU, *see* power of unfairness
- power of unfairness, 79
- referral, 117
- referral chain, 117
 - reversed, 125
- referral credibility, 116
- referrer, 118
- testimony, 5, 23
- trust, 4, 12
 - direct trust, 65
 - overall trust, 65
- trust information, 4, 23
- trust model, 3
 - cognitive, 21
 - computational, 21
- trustee, 14, 22, 44
- truster, 14, 22, 44
- unfair testimony, 5, 15
 - collusive, 93
 - endogenous method, 26
 - exogenous method, 26
 - individual, 92
- willingness, 52, 90
- witness, 45, 118
- witness profile, 67