ALLEVIATING USER BIAS IN ONLINE TRUST SYSTEMS

Tanja Pavleska

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia, March 2016

Evaluation Board:

Asst. Prof. Tanja Arh, Chair, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana

Asst. Prof. Rok Bojanc, Member, University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies, Glagoljaška 8, SI-6000 Koper
Asst. Prof. Tomaž Klobučar, Member, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL



Tanja Pavleska

ALLEVIATING USER BIAS IN ONLINE TRUST SYSTEMS

Doctoral Dissertation

ZMANJŠANJE VPLIVA PRISTRANOSTI UPORABNIKOV V RAČUNALNIŠKIH SISTEMIH ZAUPANJA

Doktorska disertacija

Supervisor: Prof. Dr. Borka Jerman Blažič

Ljubljana, Slovenia, March 2016

Index

Abstract ..... **VIII**

Povzetek ..... **IX**

Abbreviations ..... **X**

1 Introduction **1**
1.1 Research Contributions 3

2 Theoretical Background of Trust and Trust Systems **6**
2.1 Introduction 6
2.2 Basic Concepts and Related Work 6
2.2.1 Trust and Reputation 6
2.3 Related Work 7
2.4 Theories of Social Interactions Developed by Social Sciences 8
2.4.1 Social Network Analysis in Mathematical Sociology 8
2.4.2 Balance Theory 10
2.4.3 Prospect Theory and Choice Architecture 12
2.5 Theories of Social Interactions Developed by Computational Sciences 13
2.5.1 (General) Systems Theory 14
2.5.2 Belief Theory and Subjective Logic 14
2.5.2.1 Trust Reasoning with Subjective Logic 17
2.5.2.2 Transitive Trust Reasoning with Subjective Logic 18
2.5.3 Social Network Analysis in Network Theory 21
2.6 Conclusion and Discussion 22

3 The Upgraded Taxonomy of Trust Systems **24**
3.1 Introduction 24
3.2 Basic Concepts and Related Work 25
3.2.1 Systems Thinking: General Systems Theory (GST) 25
3.2.2 Case Study Research 26
3.3 Identifying Missing Elements in the Trust Taxonomy through Jordan's (GST) Taxonomy 27
3.3.1 Rate of Change: Dynamicity (D) 28
3.3.2 Purpose: Environmental Orientation (E) 28
3.3.3 Connectivity: Dense Connectivity (C) 29
3.3.4 The Proposed Taxonomy of Trust Systems: Redefining the Impact Factors 30
3.3.4.1 Active Entity Behaviour (AEB) 30
3.3.4.2 Context 32
3.3.4.3 Resilience and Evolutionism 34
3.3.4.4 Time and System Dynamics 35
3.4 Case Studies through the New Framework 36

3.4.1 BarterCast 38
3.4.1.1 Proof of concept 39
3.4.2 Yahoo! Answers 43
3.4.2.1 Proof of concept 44
3.4.3 Summary and Discussion 47
3.5 Conclusion and Future Work 49

4 User Bias Identification in Online Trust Systems 50
4.1 Introduction 50
4.2 Basic Concepts and Related Work 50
4.2.1 Online Representatives of Trust 50
4.2.2 Related Work 51
4.2.3 Prospect Theory 53
4.3 Choice Architecture 54
4.3.1 Reducing Choice Overload 54
4.3.2 Defaults 55
4.3.3 Choice over Time 55
4.3.4 Partitioning Options and Attributes 56
4.3.5 Avoiding Attribute Overload 56
4.3.6 Translating Attributes 56
4.4 Framework for Detecting and Identifying User Bias in Trust Systems 56
4.4.1 Empirical Study and Findings 57
4.4.2 Distinction Bias 58
4.4.3 Numerical-Descriptive Discrepancy and Positivity Bias 60
4.4.4 Positivity Bias and the Framing Effect 61
4.4.5 Closing the Loop: The market of Lemons on the Market of Opinions? 62
4.5 Discussion 63
4.6 Choice Architecture of Online Trust Systems 64
4.6.1 Reducing Choice Overload 64
4.6.2 Defaults 65
4.6.3 Choice over Time 66
4.6.4 Partitioning Options and Attributes 67
4.6.5 Avoiding Attribute Overload 67
4.6.6 Translating Attributes 68
4.7 Concluding Remarks and the Way Forward 68

5 A Formal Model for Alleviating User Bias in Computational Trust Systems 70
5.1 Introduction 70
5.2 Statistical analysis 70
5.2.1 Data sets description and statistics 70
5.2.2 Network Theory Analysis of the Data Sets 71
5.2.2.1 Triadic Closure and Transitivity 72
5.2.2.2 Network Motifs 73
5.2.3 Balance Theory 75
5.2.3.1 Implications of Neutrality 77
5.2.3.2 Knowledge about the Network and the Observer's Problem 79
5.2.3.3 Opinion Dependence and Conflict 80
5.2.4 Subjective Logic 81
5.2.4.1 Opinion and Sentiment Representation 81
5.2.4.2 Subjective Logic as a Predictive Model 83

5.3 The Formal Model for Bias Reduction 85
5.3.1 Extension of Transitive Trust Fusion 85
5.3.2 Evaluation and Analysis 87
5.3.2.1 Knowledge of the Network and the Observer's Problem 90
5.3.2.2 Opinion Dependence and Conflict 90
5.3.3 Extensions of the Bias Reduction Method 92
5.3.3.1 Applying Evolutionary Principles 92
5.3.3.2 Trust Transitivity and Conditional Belief Reasoning 93
5.4 Conclusion and Discussion 93

6 Concluding Remarks 96
6.1 Summary of Contributions and Results 96
6.2 Conclusion 97
6.3 Future Work 99

7 Acknowledgements 101

8 References 103

9 Appendix 118
9.1 Chapter 4: User Bias Identification in Online Trust Systems 118
Survey Questions (Reviews) 118
9.2 Chapter 5: A Formal Model for Alleviating User Bias in Computational Trust Systems 119
9.2.1 Algorithm 1: The Pseudo-code for the Extraction of Transitive Triads from the Analysed Data Sets 119
9.2.2 Balance Theory (neutral opinion analysis of the Wikipedia data set) 120
9.2.3 Subjective Logic: Sensitivity tests (if no neutral opinions are taken into account) – Slashdot Data Set 121

Bibliography 123

Biography 125

Abstract

The functionality of the Internet has outgrown its initial goal to simply transfer information between distant sites. Influenced by phenomena from the traditional world, it largely cherishes the human desire to be part of a worldwide network. It is clear that networking among people dates from long before the commercial expansion of the Internet, but there is a new perception of *socializing* and *business making* created by the impression management of people and services on the Web. The advent of social networks opens up myriad opportunities for merging the social character of trust with the technical possibilities offered by information and communication technologies. The growth of user-generated content requires social approaches for information filtering and aggregation.

The success of trust-based methods depends on the research aimed at identifying when and in what manner these methods offer most benefit. This work is a contribution to that purpose. We believe that such analysis may result in online networks providing reliable means for building trust and reputation on the Web. Moreover, the thesis aims to testify that trust awareness is needed on the Internet as it is in the real world.

This work determines the systemic features of trust and introduces a novel framework of design properties based on the principles of systems theory. We regard trust systems as human-centric, not merely technical (computational), and address some of the organismic properties that current schemes for modelling trust systems disregard. For that purpose, we explore and exploit the role of transitivity, subjectivity and user bias in trust systems and investigate their implications for the design of the computational systems that cherish the digital analogue of social trust. Through case studies, semantic, empirical and statistical analysis we identify several issues with the existing approaches that tackle trust-related problems, whereas through models based on subjective logic, systems science and balance theory we propose ways to address these problems in a multidisciplinary manner.

More concretely, this thesis demonstrates the significant discrepancy between the aim of implementation of the online trust systems and the users' perception of those systems. By employing systemic analysis, we identify four additional factors that contribute for the completeness of a novel framework for trust systems design. Then, by performing empirical study on an existing system, we demonstrate the presence of several concrete user biases that stem from, but also contribute to the discrepancy between the trust system implementation and the users' perception and behaviour. Understanding this relation by the system designers is important for reducing the complexity and improving the user-experience and the system-performance. The results and the analysis are then used to propose a methodology for user bias identification and mitigation in the form of a choice architecture for trust systems and a formal model based on subjective logic. Several extensions to the model are also proposed and placed in the plan for future work.

Povzetek

Uporaba interneta je presegla prvotni namen gole izmenjave podatkov med oddaljenimi uporabniki. Internet je omogočil uporabnikom, da ustvarjajo globalno, med seboj povezano omrežje. Njegova uporaba je ustaljene oblike povezovanja med ljudmi nadgradila v nove in širše vidike družbenih ter tržnih odnosov. Nastala družbena omrežja so ponudila priložnost nadgradnje tehničnih komunikacijsko-informacijskih tehnologij z družbeno obarvanimi koncepti zaupanja med uporabniki. Nagel porast vsebine, ki jo ustvarjajo uporabniki, zahteva družbeno podprte pristope k flitriranju in združevanju informacij.

Uspešnost pristopov zagotavljanja zaupanja je odvisna od raziskav, kdaj in na kakšen način ti pristopi zagotavljajo najboljši izkoristek. Pričujoče delo je prispevek k tem raziskavam. Verjamemo, da bo tovrstna analiza pripomogla k bolj zanesljivemu zagotavljanju zaupanja in ugleda v svetovnem spletu. Poleg tega je disertacija namenjena potrditvi teze, da je zaupanje potrebno na internetu prav tako kot v dejanskem svetu.

Predstavljeno delo določa sistemske značilnosti zaupanja in s pomočjo sistemske teorije uvede novo zasnovo lastnosti zaupanja v načrtovanje sistemov. Zaupanje obravnavamo ne samo s tehničnega, temveč tudi s človeškega stališča, in s tem pokrijemo pomankljivosti obstoječih shem modeliranja zaupanja. V delu preučimo tranzitivnost, subjektivnost in pristranost uporabnika ter vpliv teh dejavnikov na zagotavljanje zaupanja v informacijskih sistemih. Pomanjkljivosti obstoječih pristopov izpostavimo skozi študij primerov uporabe ter semantično, izkustveno in statistično analizo. Za odpravo teh pomanjkljivosti predlagamo multidisciplinarni pristop, ki združuje modele subjektivne logike, sistemsko teorijo in teorijo kognitivnega ravnovesja.

V bolj konkretnem smislu to doktorsko delo izpostavlja opazni razkorak med namenom uporabe sistemov zaupanja na spletu in načinom, kako uporabniki te sisteme dojemajo. Z uporabo sistemske analize ugotavljamo štiri dodatne dejavnike, ki prispevajo k celovitosti novega okvira zasnove sistemov zaupanja. Nato z izvajanjem empirične študije na obstoječem sistemu pri uporabnikih prikažemo obstoj več konkretnih predsodkov, ki izvirajo iz razkoraka med uporabo sistema zaupanja ter dojemanjem in obnašanjem uporabnika, hkrati pa tudi prispevajo k temu razkoraku. Razumevanje tega odnosa je pri snovalcih sistemov pomembno zaradi zmanjševanja zapletenosti ter izboljšanja uporabniške izkušnje in delovanja sistema. Rezultati in analiza so nato uporabljeni kot podlaga za predlog metodologije za ugotavljanje predsodkov uporabnikov in blaženje v obliki izbrane arhitekture za sisteme zaupanja ter formalni model na podlagi subjektivne logike. V načrtu dela za prihodnost je predlaganih in umeščenih tudi več razširitev modela.

Abbreviations

AEB	=	active entity behaviour
BC	=	BarterCast
BMA	=	belief mass assignment
BS	=	base-rate sensitivity
BT	=	balance theory
FPR	=	false positive rate
GST	=	general systems theory
MCC	=	Matthews correlation coefficient
NFC	=	need for closure
NFCC	=	need for cognitive closure
OBF	=	opposite belief favouring
RM	=	reputation mechanism
ROC	=	receiver operating characteristic
SL	=	subjective logic
SSM	=	soft systems methodology
TNA SL	=	trust network analysis with subjective logic
TPR	=	true positive rate
TSP	=	triad significance profile
UF	=	uncertainty favouring
Y!A	=	Yahoo! Answers

1 Introduction

The technological evolution of the Internet and its role as a creator of new interaction norms brought about many challenges for information and communication technologies, but also for the users of those technologies. The latest advances in virtual social platforms present additional requirements for the quality of online interactions. These requirements include reliability, trustworthiness, security and privacy, to name but a few.

Any entity (object or subject) with the ability to interact or be a part of the interactions among other entities is expected to satisfy basic security requirements. However, ensuring implementation of soft properties like trust, reputation, rating and recommendation is yet to gain adequate attention from the relevant disciplines and stakeholders.

In general, **trust** has been recognized as the crucial "lubricant" for social interactions in both the traditional and the technical world (Castelfranchi and Falcone, 2010; Fukuyama, 1996). In technical terms, trust is defined as the willingness to depend on others' actions despite the possibility of negative consequences (Josang, Ismail, and Boyd, 2007). Within the framework of this thesis, **trust systems** are defined as ones that facilitate the exchange and the processing of information relevant for making a trusting decision.

Reputation, on the other hand, is a more general opinion about an entity's trustworthiness obtained by the whole community rather than by a single entity. It is based on a history of the entity's performance in a specific context. A system that facilitates the reputation calculation is called a **reputation system**.

Nowadays trust is considered to represent people's beliefs and decisions to interact with the entities in the virtual world. This leads to the need to develop computational models of trust. Similarly to cases in the real world, the computational models that represent the relationship between two entities in a computational setting can easily fail to capture important properties of their interactions, such as how much the model resembles the reality, how close it comes to human perceptions and actions, and how much the two differ. The behaviour patterns resulting from the interactions of many trusting entities in trust systems are often more complex than the performance of each of the individuals separately. Thus, simple rules of trusting behaviour give rise to complex, emergent patterns. A major reason for these emergent properties being neither successfully captured nor adequately treated by the current formal models is the practice of addressing technical issues in a mechanistic manner. This implies considering the system merely as a sum of its components and neglecting the interactions between those components. Only a few studies can be found in the literature that try to define a framework for trust systems design (Marti and Garciamolina, 2006; Huynh, 2006; (Alani et al., 2004); Josang et al., 2007). However, none of those studies considers the interdependences between the system components in a more systemic and holistic manner. As a consequence, a framework for systemic design is required to guide the development of functioning principles of the computational trust mechanisms.

In the research literature, trust is analysed from two perspectives: cognitive and affective. The former is represented by concepts like rational choices, learning loops and pattern detection. Affective aspects, on the other hand, take into consideration the emotional drivers that accompany the trusting decisions. They are often considered to be

the ones that are responsible for the presence of so-called cognitive biases. **Cognitive bias** describes a replicable pattern in perceptual distortion, inaccurate judgment or illogical interpretation (Ariely, 2008; Kahneman, 2003; Tversky and Kahneman, 1974). It is a systematic expression of deviation from the benchmark of predictable and rational behaviour.

Despite the crucial role that biased phenomena have in the development of computational models of trust, the interdependencies between trust, information aggregation, user bias and network structure are still an unexplored avenue. Moreover, the blurred spots of where the human factor influences the trust-related processes usually lead to neglecting the impact of cognitive bias as a factor that affects the trust system design. More insights are required into how biases affect the system processes and how they manifest and impact on the information flow.

The "misfortune" related to the social trust phenomena in computational (online) systems can be formulated as:

- trust is generally too complex to even be intuitively defined, and especially to be precisely formalized or statistically analysed; and
- when it comes to representing its properties through various technicalities and formalities, they are greatly simplified for the sake of practical feasibility.

Online trust systems exist as technical implementations of traditional trusting interactions. Establishing the interdependence between these systems and the human factor is essential for reducing and understanding the inherent complexity of online social platforms, and for improving the user experience and system performance. While most computational trust models aim to truthfully detect trustworthy individuals, much less attention in the current studies is paid to how these models are perceived by the end-users, who are the core of the online trust systems. Understanding the relation between computational trust models and users' perception of those models is expected to contribute to the reduction of their complexity, while improving the user experience and the system performance.

All these concepts are discussed in the literature, which addresses social issues in the online world: similarity, trustworthiness, bias, opinion diversity and dependence; they are in reality vague and formally intangible, and the information regarding them requires aggregation in a technically meaningful manner.

Processing information of a non-deterministic nature has been the core of the artificial intelligence and information fusion areas of research. Despite the great advances in the fields of both trust management and information fusion, the high discrepancy between users' expectations and trust system performance still remains. The majority of the current approaches represent trust information as a binary value: distrust or trust with a value of 0/1 or -1/+1 and without any other attributes. Thus, they neglect all the shades coming from belief and certainty that exist between trust and distrust. They capture only the two extremes of the state of trust in human perception and reasoning. A similar line of reasoning is also met in the social and socio-computational models that formalize the balance of relationships in communities of people. Considering the fact that both trust and reputation are online qualities of an entity, an approach is required that accounts for both the statistical and the subjective properties of human decision-making.

In addition to these internal system properties, there are significant externalities that affect the performance of trust systems. The nature of the inputs and the way they are managed across various contexts influence the results provided by the trust system. Although the choice of information sources and the aggregation of the gathered information are important lines of research, the relation between the contextual relationships and the system performance to be examined for the reliability of the trust system is also important. There is practically no work in the literature that joins the

context dependence of the trust relationships and the aggregation of the information from the trust sources interconnected through these relationships. For that purpose, it is necessary to incorporate information source dependence into the reasoning about context. As a result, formal models that integrate the properties of both the network structure of the trust system and the quality of relationships among the entities in the given context need to be provided.

In this thesis, we take the stance that trust and trust systems have their own representatives in the online environment. We recognize reputation, rating and recommendation as the online representatives of trust, and reputation, rating and recommender systems as the online representatives of trust systems. We see the elements of the three systems as ones that, despite being different, more often complement one another. However, the adequate combination of the elements for harnessing reputation, providing rating and obtaining recommendation depends on many factors. The failure to recognize those factors can lead to inconsistencies in the work of the trust system, and the platform it is a part of. The user behaviour is also largely influenced by those systems, but it influences the systems' performance as well.

As a result of the described properties, several hypotheses are addressed in this PhD thesis.

The **main hypothesis** is:

Adequately structured aggregation of the trust information gathered from sources with diverse and dependent opinions can help in alleviating and limiting the effect of user bias on the trust system.

To verify the hypothesis, the following components for the trust system design are elaborated and presented:

I) Novel approach for taxonomy of trust systems:

The current frameworks for categorization of trust systems, their subcomponents and their design principles lack important elements from a systemic perspective, presenting a requirement for a novel taxonomy of trust systems.

II) Methodologies for bias identification:

To empirically detect the presence and the type of user bias in a trust system, a methodology for bias detection and limitation is required.

III) A method for the design of trust systems with a limited amount of bias:

Dependence needs to be established between the bias and the information representing social relations. This allows the effects of bias to be quantified and affected in a desirable manner through a formal model based on transitivity, subjectivity and dependence as properties of the gathered information.

Accounting for dependence, transitivity and subjectivity as contextual parameters of the mechanisms for trust aggregation can be achieved by employing the information fusion based on the framework of subjective logic, with some necessary improvements.

1.1 Research Contributions

The thesis brings the framework of systems theory and practice into the focus of designing/operating with online human-centric systems. Moreover, it highlights the fact that systematic analysis enables the understanding of the difference between the inherent and the emergent system properties. Whereas trust itself is an emergent property, it can be seen that trust has mainly been considered as an inherent systemic feature by the existing efforts for formalization and technical implementation.

This thesis determines the systemic properties of trust systems that are lacking in the current technical solutions and incorporates them in a novel taxonomy. This taxonomy is

meant to serve as guidelines for designing online trust systems. Four influencing factors are introduced and defined for that purpose. The compliance of the new taxonomy with the generic system principles is also demonstrated. Case study analysis is employed that follows the premises of case study research established in the social science literature. Semantic analysis of the state of the art of trust systems is performed, enabling the identification of two semantic clusters to which trust research belongs. Two representative case studies of real-world systems (BarterCast and Yahoo! Answers) are chosen for evaluation of the framework in a non-random manner. This is the first study to apply fundamental social science principles from the general systems theory and case study research for the purpose of theory building, taxonomical categorization and evaluation of technical human-centric solutions.

The next contribution is related to the study of reputation, recommendation and rating systems from the perspective of the biased behaviour of the users. For that purpose, an empirical study employing the principles of prospect theory is presented that identifies several types of concrete cognitive biases that the users of trust systems exhibit in their online experiences. Based on that study, a framework is provided in the form of a choice architecture for trust systems for identifying and tackling the bias phenomena for providing good user experience, system reliability and result accuracy. The importance of this work is in the revelation of trust as a core aspect in the solutions for increased complexity and user bias.

The thesis develops models and mechanisms based on subjective logic, for aggregation of trust information in online trust systems. Through statistical analysis of large data sets from real online systems, the applicability of the proposed methods and frameworks is justified and evaluated. New insights into the management of trust information in the presence of dependent information sources are provided and, as a consequence, the state of the art is advanced in both the field of aggregation of trust information and analysis of network (community) structures through the quality of the entities' relationships.

In the process of investigating and designing the means for identifying and limiting the effects of user bias on online trust systems, the thesis thoroughly studies the application of techniques and operators from the fields of information fusion and social computing and proposes adequate solutions.

The overall content of the thesis is organized as follows. In the next chapter, *trust* and *reputation* are defined in more detail. This chapter presents the progress toward their formalization, and reviews the related work from both social and computational perspectives. Subsequent chapters outline the methodology and introduce the proposed frameworks for the design and the operation of the trust system. Practical observations and insights are elaborated throughout, enabling better granularity to be introduced in the existing research map of trust systems. Chapter 3 presents a semantic study of the trust systems' research, introducing the methodology that is used in the case studies with an evaluation of the novel framework. The two specific case studies are BarterCast, a distributed reputation mechanism, and Yahoo! Answers, as being representative of the centralized social networks. Chapter 4 focuses on the concept of user bias and its effect on the design and the work of trust systems. It offers an empirical study that clearly demonstrates the presence of specific cognitive biases in these systems. Chapter 5 introduces a formal model for bias reduction based on subjective logic with the use of statistical analysis in real-world data sets, which justifies and verifies the formal methods applied. Finally, Chapter 6 contains a constructive discussion and overview of the presented results, outlining the future research directions. In each of the chapters related works on the elaborated topics are also briefly presented.

Introduction

2 Theoretical Background of Trust and Trust Systems

2.1 Introduction

The social literature on trust and its connection to community structure is exhaustive. This chapter introduces the methodology and the theoretical background used in defining the proof of the correctness of the main hypothesis. Several theories for social interactions from both the social and the computational sciences relevant for the formal apparatus used later in the thesis are presented. The application of these theories is explained as well.

The relevance and the significance of the concepts of triadic closure, transitivity and subjectivity when reasoning about trust are also discussed in this chapter. Special attention is paid to the apparatuses employed heavily in the thesis: balance theory, which is representative of the social science theories, and subjective logic, which is part of the computational theories. To justify the methodology used, the relation between the theories for social interactions and the computational theories for trust reasoning is carefully investigated as this is relevant for providing the proof of concept of the developed frameworks and formal models for trust reasoning with limited bias.

2.2 Basic Concepts and Related Work

Next, the notions of trust and reputation are introduced and a multidisciplinary overview of the efforts for their formalization is provided. The related work in each of the research areas is presented in each of the subsections.

2.2.1 Trust and Reputation

As discussed in the Introduction, from a social point of view trust is approached from two perspectives: cognitive or rational and affective or emotional (Castelfranchi and Falcone, 2010). As feelings are heavier on energy, and thoughts are heavier on information, affective properties are often responsible for the emergence of cognitive biases (Zajonc, 1980). Trust is not exclusively subject to real-world social contexts. A trusting entity is considered to be any agent capable of resolving cognitive conflicts and making preferential choices. While such agents have been part of the research into artificial intelligence, current approaches in online trust systems design mainly consider trust as a property that is inherent to the systems. This provides the desired simplicity for tackling the cognitive aspects of trust in a technical environment. However, this is not in line with the nature of the trust phenomena in the real world. The majority of research studies represent trust information as a binary value (0 or 1), without assigning any other relevant attributes (such as certainty, belief degree or provenance) to it (Klos and La Poutré, 2006). They neglect all the shades of belief and certainty that exist between trust and distrust and capture only the two extreme states of trust. Other approaches do account for the possibility of representing trust in a multi-valued manner. However, they either assign maximum certainty to each of the probability values (Abdul-Rahman and Hailes, 1997; Marti et al., 2000; Xiong and Liu, 2002; Lin, Lu, Yu, and Tai, 2005) or they do not

account for the interdependence among the trust system components and their contextual traits (Mui et al., 2001; Zhou and Hwang, 2007; Ries, 2007; Wang and Singh, 2010).

The nature of human influence often leads to neglecting the affective aspect when defining the computational analogue of trust. This, in turn, hardens the predictability of trust system behaviour. In other words, the non-linearity of social phenomena prevents a system design based on the elegant principles of mathematical linearity and probabilistic averaging. Hence, knowing the properties of the system components is insufficient to infer the properties of the system as a whole. Therefore, some working definitions must be given. Following Gambetta (1988), the following definitions are given:

Definition 1a. Trust is the subjective probability that an entity will behave in an expected manner that is beneficial, or at least not harmful.

Considering trust only as a subjective probability leaves out risk as an important concept related to trust. This fact has catalysed a vigorous debate between economists and social psychologists (Gambetta, 1988; Castelfranchi and Falcone, 2010). To consider risk by the definition of trust, the definition provided by Josang et al. (2007) is introduced:

Definition 1b. Trust is the extent of willingness to depend on others' decisions and actions, accepting the risk of undesired outcome.

Despite the generally interchangeable use of trust and reputation, a difference must be made between the two. The role of reputation mechanisms is still recognized as that of technical facilitators for managing trust.

Definition 2. Reputation is the general standing of the community regarding an entity's trustworthiness, based on the past behaviour, performance or quality of service of that entity, in a specific context. A system that facilitates reputation calculation and trust management is called a "reputation system" (or a "reputation mechanism").

Clearly, an approach that tackles trust-related issues must consider a trusting entity to be an agent that is capable of resolving cognitive conflicts and making preferential choices of an affective nature.

2.3 Related Work

Much of the research on trust decisions is rooted in game theory, where concepts like *risk* and *utility* are formally defined (Chin, 2009). Fundamental trust problems are captured by the prisoner's dilemma (Fudenberg and Tirole, 1991), a principle that demonstrates the trade-offs people make between their own benefit and the overall outcome of a game. The prisoner's dilemma is also part of the BitTorrent tit-for-tat protocol. The aim is to foster the contributions of the participating entities in the online environment (Cohen, 2003). In that context, Ba (2001) claims that people's sense of belonging to a community contributes to positive outcomes that improve community welfare. If members are responsible for their actions, a greater pressure to adhere to the rules is present. Furthermore, Fehr and Gächter (2000) have shown that individuals in that game vigorously punish selfishness, even if inducing punishment is costly to them. Systemically, this implies that control over system entities can be distributed.

In order to serve its purpose and be useful in an online setting, the trust in an entity should be part of the more general representation of the community's opinion. Therefore, in this thesis, reputation is used to denote the general standing of the community regarding an entity's behaviour. Employing reputation mechanisms (henceforth denoted as RMs) to account for an entity's behaviour was investigated long ago. It was applied in contexts dealing with information asymmetry and sanctioning undesired behaviour. In addition, such mechanisms can signal the behavioural type of unknown entities, facilitating new transactions in the system. This is known as their *signalling role*. By

analysing the effects of social and trading reputation on market maintenance, Akerlof (1970) demonstrated that low-quality goods squeeze out those of high quality due to the information asymmetry in the buyers' and sellers' knowledge in cases where reputation systems are absent. This is known as the "lemon market" problem, where "lemons" is used to denote the low-quality goods.

Despite the early work on trust relations and conflict resolution, the efforts to formalize trust came relatively late (Marsh, 1994). Marsh's work recognized the tendency of entities to group into robust clusters with similar trustworthiness. However, no distinction between groups' and individuals' properties was made, and groups were treated as entities that resolve trust issues in the same way as individuals. Scenarios where it may be undesirable to nourish trust, as in *adverse collusion*, were not considered. Despite the studies on the signalling role of reputation systems and the efforts invested by computer science (Dellarocas, 2003), the appearance of the Internet presented additional requirements for these ideas to be employed in online services as well. In that context, it was important to determine the system's organizing principles, the factors that influence its design and the interdependencies between the trusting entities. The design of online trust systems is usually influenced by factors that are both external and internal to the system. Marti and Garciamolina (2006) disassemble the trust system into three major subsystems – information gathering, scoring and ranking, and taking action – and detect three factors that influence system design: users, architecture and adversaries. However, like the majority of other known approaches to trust system design, Marti and Garciamolina consider users to be either malicious (intentionally misbehaving) or rational and strategic (intentionally well-behaving or opportunistic) when providing information or acting upon it. The process of information aggregation is the core process that embraces the effects from both the user influence (ratings and recommendations) and the technical system properties (aggregation and interpretation). In that context, in the following part of this chapter the underlying theories used for the design of trust systems and for confirmation of the hypothesis are presented.

2.4 Theories of Social Interactions Developed by Social Sciences

There is a huge body of knowledge on trust and trusting decisions collected by social sciences research. In fact, trust was initially a subject of research only in the area of social sciences. The computational research community started to consider ideas about computational trust after Marsh's work on distributed artificial intelligence (Marsh, 1994). The next section presents some of the most important works and concepts from the social sciences research on trust, reputation and similar concepts that try to explain the social interactions among community members.

2.4.1 Social Network Analysis in Mathematical Sociology

Social network analysis, which is part of the computer science disciplines, has its theoretical roots in the works of the early sociologists Georg Simmel and Émile Durkheim at the beginning of the twentieth century. It basically applies network theory to study and analyse the social interconnections among society's entities (Wikipedia, 2014). The analysis performed has shown that certain patterns arise from the relationships connecting social actors. These are the key enablers for gaining insights into the formation of societies and the phenomena that accompany it. The diversity of relationships that can be established between the social actors in a network were employed in the analysis of social networks in: anthropology, biology, economics, geography, history, organizational studies, political science, social psychology and

linguistics. Nowadays it represents a separate field in information and communication science belonging to network science.

At its root, social network analysis employs the principles of mathematics, more specifically graph theory, which is why a whole branch of sociology known as mathematical sociology was formed. In this thesis, we use a number of metrics from social network analysis. Some of them are listed below.

Homophily is a metric that specifies the extent to which actors form ties with similar versus dissimilar actors. Similarity can be defined by gender, race, age, occupation, educational achievement, status, values or any other salient characteristic (McPherson et al., 2001).

Reciprocity (or mutuality) is the extent to which two actors reciprocate each other's friendship or other interaction (Axelrod and Hamilton, 1981). A **dyad** and a **triad** are social relationships between two and three individuals, respectively.

A **bridge** is a type of social tie that connects two different groups in a social network.

Network closure is a measure expressing the completeness of relational triads. An individual's assumption of network closure (i.e. that their friends' friends are also friends) is called transitivity. According to Kruglanski and Webster (1996), transitivity results from certain traits of need for cognitive closure. **Closure** or **need for closure** (NFC) also known as **need for cognitive closure** (NFCC)) are psychological terms that describe an individual's aversion toward ambiguity, i.e., their desire for a firm answer to a question. The need for closure thus represents the will for looking for a definite answer in a context of ambiguity. This motivation is enhanced by the perceived utility from the closure (for e.g. better ability to predict the world), thus acting as an incentive for taking action. Kruglanski and Webster (1996) argue that the need for closure is mainly due to two general tendencies that people exhibit in their social interactions: the urgency tendency (to attain closure as quickly as possible) and the permanence tendency (to maintain it for as long as possible). Together, these two tendencies may contribute to reducing the extent of information processing and hypothesis generation and introducing biases in thinking. The study shows that diverse needs for closure effects appear in fundamental social psychological phenomena, including impression formation, stereotyping, attribution, persuasion, group decision-making and language use in inter-group contexts. In another study on NFC, Flynn, Reagans and Guillory (2010) found that people who had a high need for closure were more likely to assume their social contacts were connected to each other (i.e. transitivity) when this was not the case.

Ideally, people should attempt to acquire new knowledge to answer questions regarding newly arisen issues. However, because of the urgency and permanence tendencies that lie at the core of the closure process, individuals (or groups) may be compelled, consciously or unconsciously, to obtain information prematurely and irrespectively of content. Thus, a high need for cognitive closure may cause bias in: selecting the most relevant information, performing cognitive manipulations required to achieve particular outcomes, making judgments and assessments of information, or weighing up the information during the course of decision-making.

The study of NFC has shown it has implications for consumer buying behaviour (Choi, Koo, Choi and Auh, 2008). Each of the four cognitive processes are included in the processes of trusting interactions between community members in both the traditional and the virtual societies. For that reason, the biases that appear as a result of the need for cognitive closure in the system are introduced in the studies of the following chapters.

By establishing the dependence between bias and the information representing the social relations in the system, the principles of social network analysis help in the quantification of the effects of user bias on users' trusting behaviour and offer a

possibility to employ transitivity, subjectivity and dependence as properties of the gathered information. This enables the quality of the various types of relationships (positive, negative, neutral) distributed across the trust network in a context-dependent manner to be shown.

2.4.2 Balance Theory

Structural balance theory (or simply balance theory) is based on four major postulates: "the friend of my friend is my friend", "the enemy of my friend is my enemy," "the friend of my enemy is my enemy" and (in its weaker representation) "the enemy of my enemy is my friend" (Heider, 1982). A triple containing an even number of negative ties (known as "edges" in graph theory) is defined as balanced, while a triple with an odd number of negative ties is imbalanced (see Figure 1) (Heider, 1982). The presence of imbalanced triads creates psychological discomfort, i.e. "tension" for an entity that is part of the triad. Therefore, the entity aims to lower that tension by calibrating its mindset to the triad. This might also mean abandoning the triad or moving to another triad that can be adjusted so that it becomes balanced.



Figure 1: *a) Example of balanced triads* *b) Example of unbalanced triads.*

However, Heider (1982) stated: "The discussion centres on the *person* as the basic unit to be investigated." In other words, the two-person group is not the focus of attention in balance theory. This is often forgotten in the analysis of social interaction theories by computer science research. The results obtained from employing the principles from Heider's work are mistakenly considered to scale proportionally with the number of entities enrolled in the investigated interactions.

Triadic closure is defined as the property among three nodes, A, B and C, such that if a strong tie exists between A and B and A and C, there is a weak or strong tie between B and C. Tie strength is defined by the linear combination of time, emotional intensity, intimacy and reciprocity (i.e. mutuality) (Granovetter, 1983). Strong ties are associated with homophily, propinquity[1] and transitivity, while weak ties are associated with bridges. Strong ties constitute a basis of trust that can reduce resistance and provide comfort in the face of uncertainty (Krackhardt, 1992). Since people resist change and are uncomfortable with uncertainty, it is often argued that change is not facilitated by weak ties, but rather by a particular type of strong tie. Krackhardt (1992) calls these types of ties *philos*, and defines a philos relationship as "*one that meets the following three necessary and sufficient conditions:*

1. *Interaction. For A and B to be philos, A and B must interact with each other. The implication of this component is that there will be a high probability that each will have access to information that the other has, since such frequent interactions will provide opportunities to exchange such information.*
2. *Affection. For A to be philos of B, A must like B, A must feel affection for B. This evaluative component allows much of the important balance predictions of Heider and transitive closure predictions of Granovetter to hold. Heider also made*

[1] Propinquity is defined as the tendency for actors to have more ties with geographically close others.

predictions about the symmetry of the 'liking' relationship, and may assume that in most cases such relationships are symmetrical. However, one can imagine occasions when affection is not reciprocated, resulting in an asymmetric relationship.

3. *Time. A and B, to be philos, must have a history of interactions that have lasted over an extended period of time. That is, there is no such thing as instant philos. One implication here is that philos relationships cannot be studied in laboratory experiments. While one can induce short-term affective states and study 'liking', the study of philos is relegated to the field, where relationships have sufficient time to develop.*"

The property of triadic closure is too extreme to be assumed as valid across very large and complex networks. It is often considered a useful simplification of the reality and is used to understand and predict the formation and the evolution of social networks**.** However, as proved later in the thesis, the property of predictability based on triadic closure does not necessarily hold in a human-centric system. This is not surprising, given that it is hard for philos relationships (those satisfying 1, 2 and 3) to arise in a virtual context. Nevertheless, this property is often implicitly assumed by most of the computer science approaches that employ social theories in their models.

In a small-scale study of data provided from real-world networks, Doreian and Krackhardt (2001) analyse the pre-transitive balance theory mechanisms in (signed) social interactions. They realize that balance theory is not the universal mechanism for describing the formation and evolution of human relationships. As they claim:

"*…we do not know how to formally specify the coupling of a structural balance mechanism with a joint recognition of attributes mechanism. Clearly, structural balance cannot be the entire story and our results suggest it may not be the dominant story […] whatever mechanisms are operating, they generate inequality in the receipt of negative ties.*"

However, although the data in this study are longitudinal, the longevity of relationships is not discussed. The authors speak of the "first" tie {i,j}, but it is impossible in a virtual network to know which tie was formed first in the triple, and whether the other two connections ({j,k} and {i,k}) were aware of both the presence of the tie {i,k} and of its sign. The study shows that individual actors are not the sole basis on which positive and negative ties are determined – the "people are judged by the company they keep" principle was proven to have a great role in the overall outcome. Furthermore, it was found that there can be external influences as well (for instance, from other people outside the network, from the media, etc.). Finally, there are many other subtleties related to people's (i.e. users') sentiments that cannot always be captured by the formal computational mechanisms, such as the difference between taste and preference (Ariely, 2008; Ariely, Loewenstein and Prelec, 2006), trust and friendship (Adamic, Lauterbach, Teng and Ackerman, 2011), etc.

One study that includes balance theory in social network analysis was provided by Leskovec, Huttenlocher and Kleinberg (2010). It should be noted that here, context and link semantics were not related to the theories investigated by Leskovec et al (2010). The analysis of cyclic instead of transitive triads in the analysis does not provide useful information regarding the applicability of balance theory in this context.

Balance theory enables the establishment of the connection between user bias as a design and operation issue in trust systems and the quality of the social interactions in these systems. The insights from this theory are integrated in the taxonomy presented in the following chapter, but the main use of balance theory application is found in the development of formal methods for bias identification and reduction in the later chapters.

2.4.3 Prospect Theory and Choice Architecture

The main goals in designing online trust systems are to facilitate online interactions and to help people make decisions. In this regard, trust systems come close to other systems designed to aid decision-making in various contexts.

Prospect theory is a theory from behavioural economy that was first introduced by the two psychologists Kahneman and Tversky (Kahneman and Tversky, 1979). They introduced it as a critique of the utility theory, which is a descriptive model of decision-making that neglects some essential aspects of human nature. It uses "utility" to describe enjoyment and content in people, viewing them as entities that always aim to maximize their own utility. However, research on prospect theory has found that people do not process information in such a rational way. Just as they are not aware of the engagement of the brain for translating vision into sight, they are not aware of the kinds of computations the brain makes in editing and evaluating choices. People make decisions according to how their brains process and understand information and not solely on the basis of the inherent utility that a certain option possesses for a decision-maker. Therefore, Kahneman and Tversky developed prospect theory as an alternative model for decision-making. They found empirically that people underweight outcomes that are merely probable compared with outcomes that are obtained with certainty. Furthermore, they demonstrated that people generally discard components that are shared by all prospects under consideration. One of the most exploited results from prospect theory is the one demonstrating that people's attitudes toward risks concerning gains often greatly differ from their attitudes toward risks concerning losses. It is important to note here that in prospect theory, value is assigned to gains and losses rather than to final assets, whereas probabilities are replaced by decision weights.

Prospect theory is descriptive and empirical in nature and it is designed to explain a common pattern of choice. It looks at two parts of the decision-making process: the editing, or framing, phase and the evaluation phase. The editing phase encompasses the framing effects, which refer to the way in which a choice or an option can be affected by the order or manner in which it is presented to a decision-maker. The paradox here is that framing effects are often embedded in decision problems in such a way that the decision-makers rarely realize the disproportionate impact that these (framing) effects have on them. Thus, they frequently remain unaware of them and make decisions based on intuitive assessments of predetermined options. Therefore, any descriptively adequate model of choice must be sensitive to these framing effects. The evaluation phase involves the decision process of choosing among options; this decision is influenced by two processes, one related to subjective value, the other to perceptual likelihood. As a result of these effects of psychological, social, cognitive and emotional factors on individuals' decisions, tastes and preferences, a "wealth" of biases have been discovered and studied by the proponents of prospect theory. Another line of research from behavioural economy that embraces and resembles such ideas is found in the works of Ariely (2008), Ariely, Huber and Wertenbroch (2005) and Ariely et al. (2006). Although prospect theorists have identified important human decision-making biases, they have remained silent on the question of the origin of these biases. McDermott, Fowler and Smirnov (2008) have created a model that points to the possibility that the preferences which are in line with prospect theory have origin in evolutionary psychology. They have derived a risk-sensitive model based on optimal foraging theory to generate an explanation of the origin and function of context-dependent risk aversion and risk-seeking behaviour.

Choice architecture is a framework that considers the design of different ways in which choices can be presented to users/consumers and the impact of that design on the

users' decision-making. The features adopted by this framework include the number of choices presented to the user, the description of the given attributes and the way in which a "default" can influence users' choice.

The concept was coined by Richard Thaler and Cass Sunstein in their book *Nudge: Improving Decisions about Health, Wealth, and Happiness* (Sunstein and Thaler, 2009). A concrete design of choice architecture was pointed out in their work as a way to improve consumer decision-making by limiting the impact of biases and errors arising from people's bounded rationality. The approach they proposed consists of "nudging" individuals toward choices that are in their best interest without actually limiting their choice. It is based on the idea that it is both possible and legitimate for private and public institutions to affect behaviour while also respecting freedom of choice. This is also known as "libertarian paternalism" (Thaler and Sunstein, 2003).

Initially, three features were mentioned as key features of choice architecture: default, giving feedback and expecting error. More recently, other features have also been added, of which the most notable are: incentives, mappings and structuring complex ideas.

Default is what happens if nothing is done, such as leaving a computer unused until the screen saver appears. The main point with default options is that whatever is chosen as the default has a very good chance of being selected. Furthermore, people respond to feedback; for instance, signs/icons that denote a non-reputable user of a certain web service can aid the decision-making, but also make the web service more user-friendly. An example of expecting error can be seen in the design of the Paris metro card, which allows users to insert it into an electronic turnstile in any of four ways to gain entrance to the metro, in contrast to the tickets for exiting car parks, where there is only one correct way to insert the ticket. Google is developing choice architecture to remind Gmail users when they forget an attachment or may be about to send a rude email. If the word "attachment" is mentioned in the text of the email and the sender doesn't include an attachment, they would be prompted to do so.

The use of prospect theory and choice architecture for bias identification, empirical detection of its elements and the concrete types of user bias in trust systems is visible in the work presented in Chapter 4, where most of the parts of the decision-making process defined by these theories are applied. Further development beyond the state of the art is provided as the different options presented to people are put together or separated for evaluation. Different representations of essentially the same information are also analysed as perceived and assessed by the users of trust systems. While the motivation of these undertakings has roots in prospect theory, this work has explored the presented subjects in a computational setting, through concrete empirical studies. The results obtained testify to the particular types of biases bound to the trusting interactions in the online world.

2.5 Theories of Social Interactions Developed by Computational Sciences

The advent of virtual social networks and the great shift towards online social interactions led to an equal shift in the scientific research of the phenomena accompanying these interactions. Thus, many disciplines that study these phenomena appeared in the technical world, representing the virtual analogue of a certain discipline doing similar research for the traditional (social) world. On the one hand, this enabled interesting avenues to be discovered for research in the areas studying human phenomena that accompany the social interactions in a non-physical world. On the other hand, it created a danger of scientific reductionism, whereby the same social theories that were proved to hold in the traditional world are considered to hold as they were used for modelling and predicting the social interactions in the virtual world. One of the insights provided in this thesis is

the identification of the danger of reductionism.

This section presents some of the most important research studies from computational sciences that relate to trust research and pinpoints its relation to the specific social theories presented in the previous section. At the same time, it provides information about the formal apparatus employed later in the thesis.

2.5.1 (General) Systems Theory

Systems theory is the interdisciplinary study of systems in general, with the goal of bringing together and explaining the principles that can be applied to all types of systems at all levels in all fields of research. The term does not have a precise meaning, but systems theory can reasonably be considered a specialization of *systems thinking*. The term originates from Bertalanffy's general system theory (GST) (Bertalanffy, 1969), but is used and applied across many other fields, such as the action theory of Talcott Parsons (Burger, Parsons and Toby, 1978) and the social systems theory of Niklas Luhmann (1987). As a transdisciplinary, interdisciplinary and multi-perspectival domain, the area brings together principles and concepts from ontology, philosophy of science, physics, computer science, biology and engineering, as well as geography, sociology, political science, psychotherapy (within family systems therapy) and economics, among others. Systems theory thus serves as a bridge for interdisciplinary dialogue between autonomous areas of study as well as within the area of systems science itself. In systems science, it is argued that the only way to fully understand why a problem or element occurs and persists is to understand the parts in relation to the whole. Such a view of a system is referred to as a holistic view.

Although systems theory is mainly considered to have an engineering nature, it tackles issues from both the social and the computational sciences. Many of the concepts and principles from GST employed in this study do not have only a computational character, but are also rooted and employed in the social sciences.

This work follows the principles of systems theory. Special attention is devoted to explaining and establishing the basis for such choice in the context of trust systems in the next chapter. The design issues related to trust systems are detected and addressed, and a taxonomy is developed that serves as a framework for design guidelines and principles used in the later chapters. With this, the first hypothesis outlined in the introducing chapter is approached, demonstrating the need for a novel taxonomy of trust. The lack of such a framework for the categorization of trust systems, their subcomponents and their design principles implies lack of important elements from a systemic perspective. The postulates of systems theory are described in greater depth in the next chapter.

2.5.2 Belief Theory and Subjective Logic

In standard logic, propositions are considered to be either true or false, while in probabilistic logic the arguments are expressed as a probability in the range [0, 1]. However, a fundamental aspect of the human condition is that nobody can ever determine with absolute certainty whether a proposition is true or false, or determine the probability of something with 100% certainty. In addition, whenever the truth of a proposition is assessed, it is always done by an individual, and it can never be considered to represent a general and objective belief. This indicates that important aspects are missing in the way standard logic and probabilistic logic capture people's perception of reality, and that these reasoning models are more designed for an idealized world.

Belief theory is a framework related to probability theory, but where the values of the probabilities over the set of possible outcomes do not necessarily add up to 1, and the

remaining probability is assigned to the union of possible outcomes (Shafer, 1976). Several concepts of the Dempster-Shafer theory of evidence are recalled here in order to introduce the notations and the terminology of the formal apparatus employed in the thesis.

The first step in applying the Dempster-Shafer belief model is to define a set of possible situations, which is called the "frame of discernment". A frame of discernment delimits a set of possible states of a given system, only one of which is assumed to be true at any one time. The frame of discernment can, for example, be the set of six possible outcomes of throwing a dice, so that the outcome of a particular instance of throwing the dice becomes the state variable.

In this thesis, standard set theory is used to describe frames of discernment, but the term "state" is used instead of "set" because it is more relevant to the field of application. The elementary states in the frame of discernment Θ are called "atomic states" because they do not contain substates (i.e. one cannot extract any subsets out of an atomic state). The powerset of Θ, denoted by 2^{Θ}, contains the atomic states and all possible unions of the atomic states, including Θ.

Definition 3: (belief mass assignment) Let $\Theta = \{\theta_i \; ; \; i = 1, \cdot \cdot \cdot k\}$ be a frame of discernment – a finite set of exhaustive and exclusive possible values for a state variable of interest. A belief mass assignment (or BMA), denoted by *m*, is defined as a belief distribution function from the powerset 2^{Θ} to [0, 1] satisfying:

$$m(\emptyset) = 0 \;\; and \sum_{x \subseteq \Theta} m(x) = 1$$

The values of a BMA are called "belief masses". Each subset $x \subseteq \Theta$ such that $m(x) > 0$ is called a "focal element" of Θ.

A belief mass $m_{\Theta}(x)$ expresses the belief assigned to the state *x* and does not express any belief in substates of *x* in particular. In contrast to belief mass, the belief in a state must be interpreted as an observer's total belief that a particular state is true. With this in mind, we give the following definitions that form the basis of the belief model and the subjective probability model (Jøsang, 2001):

Definition 4: (belief function) Let Θ be a frame of discernment, and let m_{Θ} be a BMA on Θ. Then the belief function corresponding to m_{Θ} is the function b: $2^{\Theta} \rightarrow [0, 1]$ defined by:

$$b(x) = \sum_{y \subseteq x} m_{\Theta}(y), \quad x, y \in 2^{\Theta}$$

Similarly to belief, an observer's disbelief must be interpreted as the total belief that a state is not true. The following definition expresses this:

Definition 5: (disbelief function) Let Θ be a frame of discernment, and let m_{Θ} be a BMA on Θ. Then the disbelief function corresponding with m_{Θ} is the function d: $2^{\Theta} \rightarrow [0, 1]$ defined by:

$$d(x) = \sum_{y \cap x = \emptyset} m_{\Theta}(y), \quad x, y \in 2^{\Theta}$$

The following definition expresses uncertainty.

Definition 6: (uncertainty function) Let Θ be a frame of discernment, and let m_{Θ} be a BMA on Θ. Then the uncertainty function corresponding with m_{Θ} is the function u: $2^{\Theta} \rightarrow [0, 1]$ defined by:

$$u(x) = \sum_{\substack{y \cap x \neq \emptyset \\ y \not\subset x}} m_{\Theta}(y), \quad x, y \in 2^{\Theta}$$

The following additivity principle also holds:

$$b(x) + d(x) + u(x) = 1, \qquad x \in 2^{\theta}, x \neq 0$$

Definition 7: (relative atomicity) Let Θ be a frame of discernment, and let $x,y \in 2^{\Theta}$. Then for any given $y \neq 0$ the relative atomicity of x to y is the function a: $2^{\Theta} \rightarrow [0, 1]$ defined by:

$$a(x/y) = \frac{|x \cap y|}{|y|}, \quad x, y \in 2^{\Theta}, y \neq 0$$

Evidently, $x \cap y = \emptyset \Rightarrow a(x/y) = 0$ and $y \subseteq x \Rightarrow a(x/y) = 1$. In all other cases the relative atomicity will be a value between 0 and 1.

A frame of discernment with a corresponding BMA can be used to determine a probability expectation value for any given state. Uncertainty contributes to the probability expectation, but will have a different weight depending on the relative atomicities.

Belief calculus is suitable for approximate reasoning in situations of partial ignorance regarding the truth of a given proposition. The main idea behind belief theory is to abandon the additivity principle of probability theory, i.e. that the sum of probabilities on all pairwise disjoint states must add up to one. Instead, it gives observers the ability to assign so-called belief mass to the powerset of the state space. The main advantage of this approach is that ignorance, i.e. the lack of evidence about the truth of the states, can be explicitly expressed, e.g. by assigning belief mass to the whole state space.

Subjective logic is a type of probabilistic logic that allows probability values to be expressed with degrees of uncertainty (Jøsang, 2001). It represents a specific belief calculus that uses a belief metric called *opinion* to express beliefs (Jøsang, 2001). The subjective opinion model of this theory extends the traditional belief function model in the sense that opinions take base rates into account whereas belief functions ignore base rates. The representation of the uncertain probabilities in subjective logic will be based on a belief model similar to the one used in the Dempster-Shafer theory of evidence. Thus, employing belief, disbelief, uncertainty and relative atomicity, the subjective opinion is defined as follows:

Definition 8: (the opinion space) Let Θ be a binary frame of discernment with two atomic states x and $\neg x$, and let m_{Θ} be a BMA on Θ where $b(x)$, $d(x)$, $u(x)$ and $a(x)$ represent the belief, disbelief, uncertainty and relative atomicity functions on x in 2^{Θ} respectively. Then the opinion about x, denoted by ω_x, is the tuple defined by:

$$\omega_x \equiv \big(b(x), d(x), u(x), a(x)\big).$$

For clarity in the notation, the belief, disbelief, uncertainty and relative atomicity functions are denoted as b_x, d_x, u_x and a_x respectively, or simply as b, d, u and a when x is understandable by the context. Hence, an opinion denoted by $\omega_x^A = (b, d, u, a)$ expresses the relying party A's belief in the truth of statement x. When a statement, for example, says "*Party* X *is honest and reliable regarding* σ", then the opinion about the truth of that statement can be interpreted as trust in X within the scope of σ. Here b, d and u represent belief, disbelief and uncertainty respectively, where $b, d, u \in [0,1]$ and $b + d + u = 1$. The parameter $a \in [0,1]$ is also called the "base rate" (for a binary frame of discernment), and is used for computing an opinion's probability expectation value that can be determined as $E(\omega_x^A) = b + au$. More precisely, a determines how uncertainty contributes to the probability expectation value $E(\omega_x^A)$. In the absence of any specific evidence about a given party, the base rate determines the a priori trust that would be put in any member of the community.

An important property of opinions is that they are equivalent to Beta or Dirichlet

probability density functions (pdf) under a specific mapping (Jøsang and Elouedi, 2007; Jøsang, 2001). This equivalence can be derived from the distribution of belief masses. The idea of subjective logic is to extend probabilistic logic by expressing uncertainty about the probability values themselves, meaning that it is possible to reason with arguments in the presence of uncertain or incomplete evidence. Subjective logic is directly compatible with binary logic, probability calculus and classical probabilistic logic. The advantage of using subjective logic is that real-world situations can be more realistically modelled, and that the results more correctly reflect the ignorance and uncertainties that necessarily result from partially uncertain input arguments. Finally, the opinions used in subjective logic offer significantly greater expressiveness than binary or probabilistic values by explicitly including degrees of uncertainty, thereby allowing an analyst to specify "*I don't know*" or "*I'm indifferent*" as input argument.

The context-dependent aggregation mechanisms for trust reasoning introduced later in Chapter 5 are presented precisely by employing the information fusion framework of subjective logic, with some necessary improvements. Each of the following chapters adds to and elaborates on these improvements, in order to finally wrap all of them into a single formal model for bias reduction. The employment of subjective logic enables us to account for dependence, transitivity and subjectivity as contextual parameters of the mechanisms for trust aggregation defined as major elements that a system designer needs to deal with.

2.5.2.1 Trust Reasoning with Subjective Logic

Subjective logic defines a number of operators (Jøsang, 2001; Jøsang, Ažderska and Marsh, 2012; Josang, Daniel and Vannoorenberghe, 2003; Josang, 2008), where some represent generalizations of binary logic and probability calculus operators, whereas others are unique to belief theory because they depend on belief ownership. Here, the basic operators for opinion aggregation are introduced. The *discounting* and the *consensus* operators are briefly presented, and then the operators for transitive trust reasoning that will be exploited throughout the thesis chapters.

The discounting operator defined below can be used to derive trust from transitive paths, and the consensus operator can be used to derive trust from parallel paths. These operators are defined as follows Josang (2013):

Discounting is an operator used to compute the trust score derived from the transitive paths in a certain network. Assume two agents A and B, where A has referral (indirect) trust in B, denoted as $\omega_B^A = (b_B^A, d_B^A, u_B^A, a_B^A)$. Furthermore, let B have functional (direct) trust in C, denoted as $\omega_C^B = (b_C^B, d_C^B, u_C^B, a_C^B)$. A's indirect functional trust in C can then be derived by discounting B's trust in C with A's trust in B. This derived trust is denoted by $\omega_C^{A:B} = (b_C^{A:B}, d_C^{A:B}, u_C^{A:B}, a_C^{A:B})$. By using the symbol $\otimes$ to designate this operator, it can be written as $\omega_C^{A:B} = \omega_B^A \otimes \omega_C^B$.

$$\begin{cases} b_C^{A:B} = b_B^A / b_C^B \\ d_C^{A:B} = b_B^A / d_C^B \\ u_C^{A:B} = d_B^A + u_B^A + b_B^A u_C^B \\ a_C^{A:B} = a_C^B \end{cases} \quad (1)$$

Consensus is an operator used to fuse two (possibly conflicting) beliefs into one. Let $\omega_C^A = (b_C^A, d_C^A, u_C^A, a_C^A)$ and $\omega_C^B = (b_C^B, d_C^B, u_C^B, a_C^B)$ represent A's and B's trust in C, respectively. The opinion $\omega_C^{A\diamond B} = (b_C^{A\diamond B}, d_C^{A\diamond B}, u_C^{A\diamond B}, a_C^{A\diamond B})$ is called the consensus between ω_C^A and ω_C^B. The result from the consensus operator can be understood as representing the opinion, i.e. the trust that an imaginary agent cherishes for C, as if this

agent is the representative for both A and B. The symbol that is used to designate the consensus operator is $\oplus$. Thus, we can write $\omega_C^{A\diamond B} = \omega_B^A \oplus \omega_C^B$. There are cases of application of this operator:

Case I: $u_C^A + u_C^B - u_C^A u_C^B \neq 0$ *Case II*: $u_C^A + u_C^B - u_C^A u_C^B = 0$

$$\begin{cases} b_C^{A\diamond B} = \frac{b_C^A u_C^B + b_C^B u_C^A}{u_C^A + u_C^B - u_C^A u_C^B} \\ d_C^{A\diamond B} = \frac{d_C^A u_C^B + d_C^B u_C^A}{u_C^A + u_C^B - u_C^A u_C^B} \\ u_C^{A\diamond B} = \frac{u_C^A u_C^B}{u_C^A + u_C^B - u_C^A u_C^B} \\ a_C^{A\diamond B} = a_C^A \end{cases} \qquad \begin{cases} b_C^{A\diamond B} = \frac{\gamma^{A/B} b_C^A + b_C^B}{\gamma^{A/B} + 1} \\ d_C^{A\diamond B} = \frac{\gamma^{A/B} d_C^A + d_C^B}{\gamma^{A/B} + 1}, \\ u_C^{A\diamond B} = 0 \\ a_C^{A\diamond B} = a_C \end{cases} \quad (2)$$

where the relative weight $\gamma^{A/B} = \lim u_C^B / u_C^A$.

The effect of the consensus operator is to amplify belief and disbelief and reduce uncertainty, i.e. to increase the confidence in the expectation value. In cases where the subjective opinions are probability values ($u = 0$), *Case II* produces the weighted average of probabilities. The γ term is called *relative dogmatism*.

The consensus of two totally uncertain opinions results in a new totally uncertain opinion. The most important property of this operator is that it allows for two dogmatic opinions (where there is no uncertainty, i.e. $u = 0$) to be combined. This result makes the consensus operator more general than the operators defined in the Dempster-Shafer theory of evidence where the combination of totally conflicting beliefs is excluded.

A more in-depth elaboration on how opinions are formed from belief theory calculus and how each of the operators is obtained can be found in the works of Josang (2013). Here, only the essentials necessary for understanding the formal apparatus of the thesis are presented.

2.5.2.2 Transitive Trust Reasoning with Subjective Logic

Assume two agents A and B where A trusts B, and B believes that proposition *x* is true. Transitivity would then imply that agent A also trusts that the proposition *x* is true (as shown in Figure 2). Clearly, this assumes that there is some kind of mechanism for B to convey its recommendation about *x* to A. In subjective logic, trust and belief are formally expressed as opinions. The transitive linking of these two opinions consists of discounting B's opinion about *x* by A's opinion about B (using the discounting operator), in order to derive A's opinion about *x* (Jøsang, Marsh and Pope, 2006).



Figure 2: *Transitive trust fusion*.

However, as trust is a human phenomenon, trust transitivity itself is a human phenomenon. Therefore, an objective transitivity as defined in mathematics is difficult to achieve. Trust transitivity is a different (subjective) interpretation of the concept of transitivity. As a consequence, only a few operators defined in subjective logic can capture the situations that include transitive trust reasoning, and that take into account the contextual nature of trust. For instance, one of the operators accounts for a situation where A disbelieves that B is giving good advice, and so it takes the opposite of what B is

recommending as being true. The three operators that can handle various situations in transitive trust reasoning are presented as follows:

Uncertainty favouring discounting: A's disbelief in (the recommending agent) B means that A thinks B is ignorant about the truth value of the proposition *x*. As a result, A also ignores the truth value of *x*. It is important to note here that not only humans are considered a subject of trust, but also other entities, such as music items, books, products, etc.

Let us now consider two agents A and B, where A's opinion about B's recommendations is denoted as $\omega_B^A = (b_B^A, d_B^A, u_B^A, a_B^A)$. Furthermore, let B's recommendation about the proposition *x* be expressed to A as $\omega_x^B = (b_x^B, d_x^B, u_x^B, a_x^B)$. A's trust in *x* can then be derived by discounting B's opinion about *x* with A's trust in B, and the same as previously can be denoted as $\omega_x^{A:B} = (b_x^{A:B}, d_x^{A:B}, u_x^{A:B}, a_x^{A:B})$, where:

$$\begin{cases} b_x^{A:B} = b_B^A b_x^B \\ d_x^{A:B} = b_B^A d_x^B \\ u_x^{A:B} = d_B^A + u_B^A + b_B^A u_x^B \\ a_x^{A:B} = a_x^B \end{cases} \quad (3)$$

By using the symbol $\otimes$ to designate this operator, the operation can be written as $\omega_x^{A:B} = \omega_B^A \otimes \omega_x^B$.

This operator is associative but not commutative, implying that the combination of opinions can start at either end of the path, but the order in which the opinions are combined matters. In a path with more than one recommending entity, opinion independence must be assumed, which, for example, translates into not allowing the same entity to appear more than once in a transitive path. For example, let $\omega_B^A = \{b_B^A = 0.1, d_B^A = 0.8, u_B^A = 0.1)$ and $\omega_x^B = \{b_x^B = 0.8, d_x^B = 0.1, u_x^B = 0.1)$. Then $\omega_x^{A:B} = \{0.08, 0.01, 0.91)$, maximizing the uncertainty in the resulting opinion.

Opposite belief favouring discounting: A's disbelief in the recommending agent B means that A thinks that B consistently recommends the opposite of their real opinion about the truth value of *x*. As a result, A not only disbelieves in *x* to the degree that B recommends belief, but they also believe in *x* to the degree that B recommends disbelief in *x*, because the combination of two disbeliefs results in belief in this case. This operator models the principle "*your enemy's enemy is your friend*". That might be the case in some situations, and the operator should only be applied when the situation makes it plausible. It is not certain whether it is meaningful to model more than two arcs in a transitive path with this principle. In other words, it is doubtful whether the enemy of your enemy's enemy is necessarily your enemy too.

Formally, let A and B be two agents once again, where A's opinion about B's recommendations is denoted as $\omega_B^A = (b_B^A, d_B^A, u_B^A, a_B^A)$, and B's recommendation about the proposition *x* is expressed to A as $\omega_x^B = (b_x^B, d_x^B, u_x^B, a_x^B)$. A's trust in *x* can then be the same as previously derived by discounting B's opinion about *x* with A's trust in B, denoted as $\omega_x^{A:B} = (b_x^{A:B}, d_x^{A:B}, u_x^{A:B}, a_x^{A:B})$, only now:

$$\begin{cases} b_x^{A:B} = b_B^A b_x^B + d_B^A d_x^B \\ d_x^{A:B} = b_B^A d_x^B + d_B^A b_x^B \\ u_x^{A:B} = u_B^A + (b_B^A + d_B^A) u_x^B \\ a_x^{A:B} = a_x^B \end{cases} \quad (4)$$

The same symbol $\otimes$ is used to designate the operation, as it is the same operation

(discounting) in another context, and thus $\omega_x^{A:B} = \omega_B^A \otimes \omega_x^B$.

Recalling the balance theory principles of interaction, it is evident that the opposite belief favouring operator models the same principles, but generalized to subjective opinions. Thus, transitive trust reasoning can, under certain conditions, be considered a generalization of balance theory. Determining these conditions and giving contextual meaning to the interactions considered by balance theory is presented in Chapter 5.

Base rate sensitive discounting: In the transitivity operators defined above, the parameter a_B^A had no influence on the discounting of the recommended $(b_{x,}^B, d_x^B, u_x^B)$ parameters. This can seem counter-intuitive in many cases. It can be explained by the following example described by Josang (2006). Imagine a stranger coming to a town that is known for its citizens being honest. The stranger is looking for a car mechanic, and asks the first person he meets to direct him to a good car mechanic. The stranger receives the reply that there are two car mechanics in town, David and Eric, where David is cheap but does not always do quality work, and Eric might be a bit more expensive, but he always does a perfect job. Translated into the formalism of subjective logic, the stranger has no other info about the person he asks than the base rate that the citizens in the town are honest. The stranger is thus ignorant, but the expectation value of a good piece of advice is still very high. Without taking a_B^A into account, the result of the definitions above would be that the stranger is completely ignorant about which of the mechanics is the best. An intuitive approach would then be to let the expectation value of the stranger's trust in the recommender be the discounting factor for the recommended $(b_{x,}^B, d_x^B)$ parameters.

Formally, the base rate sensitive discounting of a belief $\omega_x^B = (b_x^B, d_x^B, u_x^B, a_x^B)$ by a belief $\omega_B^A = (b_B^A, d_B^A, u_B^A, a_B^A)$ produces $\omega_x^{A\diamond B} = (b_x^{A\diamond B}, d_x^{A\diamond B}, u_x^{A\diamond B}, a_x^{A\diamond B})$, where:

$$\begin{cases} b_x^{A\diamond B} = E(\omega_B^A) b_x^B \\ d_x^{A\diamond B} = E(\omega_B^A) d_x^B \\ u_x^{A\diamond B} = 1 + E(\omega_B^A) u_x^B - E(\omega_B^A) \\ a_x^{A\diamond B} = a_x^B \end{cases} \quad (5)$$

where the probability expectation value $E(\omega_B^A) = b_B^A + a_B^A u_B^A$.

However, there are certain conditions under which this operator should be applied. Josang again explains this with the example of the town of honest citizens. Let the stranger A have the opinion $\omega_B^A = \{0, 0, 1, 0.99)$ about the first person B he/she meets, i.e. the opinion has no basis in evidence other than a very high base rate defined by $a_B^A = 0.99$. If the person B now recommends to A the opinion $\omega_x^B = \{1, 0, 0, a)$, then, according to the base rate sensitive discounting operator, A will have the belief $\omega_x^{A:B} = \{0.99,\ 0, 0.01, a)$ in x. In other words, the highly certain belief $\omega_x^{A:B}$ is derived on the basis of the highly uncertain belief ω_B^A, which may seem counter-intuitive. This potential problem could be amplified as the trust path gets longer. Therefore, Josang recommends as a safety principle only performing a base-rate sensitive discounting to the last link in the transitive trust chain. This implies some loss of information, which is assumed not to be harmful if there is not a large number of such trust chains.

A very useful source for contextual analysis of information diffusion and interaction patterns that provides a description of the diffusion patterns in seven online domains ranging from communication platforms to networked games to microblogging services, each involving different types of content and modes of sharing, was presented by Goel, Watts and Goldstein (2012). They found similar patterns across all domains, but the major finding is that the vast majority of cascades are small and described by a handful of

simple tree structures that terminate within one degree of an initial adopting seed. This finding may be used to justify the applicability of subjective logic to one-hop (or very short) referral chains, which is the condition for applying base-rate sensitive discounting. Finally, even for the largest cascades observed, the authors find that the bulk of adoptions take place within one degree of a few dominant individuals. Although this speaks about the great impact that an individual entity can have, it does not point to any way of assessing, quantifying, let alone influencing the extent of this impact.

2.5.3 Social Network Analysis in Network Theory

Network theory in general concerns itself with the study of graphs as a representation of either symmetric relations or, more generally, of asymmetric relations between discrete objects. Applications of network theory include logistical networks, the World Wide Web, the Internet, gene regulatory networks, metabolic networks, social networks, epistemological networks, etc. (Newman, 2003). One of the most significant and also most popular aspects of network theory is social network analysis. Similarly to its applications in a (traditional) social context, it has also been used in computational sciences to understand the diffusion of innovations, news and rumours, or to examine the spread of diseases and thus address some health-related issues. Social network analysis has also been applied to the study of markets, to examine the role of trust in exchange relationships and of the social mechanisms in setting prices (Becker and Murphy, 2003). Furthermore, it has been used to study the influence and impact of certain political movements and social organizations on the social structures that appear on the Internet. It has also been used to conceptualize scientific disagreements as well as academic prestige.

Although the purpose of applying social network analysis discipline from network theory seems identical to the one from the social sciences described in the previous chapters, there are unique properties of network theory that are only tied to the computational character of the methodology and the principles of the discipline. The possibility of collecting, processing and analysing big data sets from real-world implementations through advanced technologies and sophisticated formal methods is a great advantage over the mainly descriptive methods offered and used by the social sciences. This implies that the whole field of network theory also employs different terminology and approaches problem solving in a conceptually different manner.

In the context of computational sciences, Adamic et al. (2012) have shown that, although stronger ties are individually more influential, the more abundant weak ties are the ones that are responsible for the propagation of novel information. This is not a surprising result as it was shown long ago that the same stands for the ties formed in the traditional social networks (Granovetter, 1983). Its significance, however, lies in the fact that it provides a means to bridge the set of contextual traits responsible for this result in the (different) worlds of virtual and traditional social networks. In another study by the same authors, they also conjecture that it pays off to target less influential individuals in a greater number when the cost of engagement is small (or negligible) in order to achieve greater diffusion (of ideas mainly) (Bakshy, Hofman, Mason, & Watts, 2011). Both of these findings, however, point to the large set of opportunities to achieve great diffusion of ideas, opinions and products. They also reveal the potential danger of targeted attacks for exploiting the benefits of these means and processes for malicious purposes: organizing mass movements, proliferating bogus content, etc.

As was elaborated previously when discussing social network analysis from the perspective of social sciences, triadic closure was made popular by Mark Granovetter (1983) who actually synthesized Heider's theory of social (or cognitive) balance with a Simmelian understanding of social networks. (Remember that cognitive balance in

general refers to the propensity of two individuals to want to feel the same way about an object. If the triad of three individuals is not closed, then the person connected to both of the individuals will want to close this triad in order to achieve closure in the relationship network.) Employing formalisms from probability and mathematics, the so-called "weak tie hypothesis" argues that if A is linked to both B and C, then there is a greater-than-chance probability that B and C are also linked to each other (Granovetter, 1983). In a trust network, triadic closure is likely to develop due to the property of transitivity. If a node A trusts node B, and node B trusts node C, node A is expected to also cherish trusting affinity towards C. In a social network, strong triadic closure occurs because there is increased opportunity for nodes A and C with common neighbour B to meet and therefore create at least weak ties. Node B also has the incentive to bring A and C together to decrease the latent stress in two separate relationships (David and Jon, 2010). Triadic closure is believed to be a good model for how networks will evolve over time.

In this thesis, the apparatus and the models offered by network theory to study the network structure and the system macro-behaviour of implemented social networks and its users are heavily exploited in Chapter 5 for the development of a methodology for reducing the effects of user bias on the functioning of trust systems.

2.6 Conclusion and Discussion

There is a trend of taking a reductionist approach in straightforwardly translating the social theories of trusting relations into the virtual world. This kind of approach, although handy at first glance, brings little or no practical value to computational trust research, to the extent that it may even be harmful if accepted without first determining the conditions and the design assumptions under which it would hold. As a study of strategic decision-making that embraces mathematical models of conflict and cooperation, game theory was and still remains the predominant theory used in economics, psychology, logic, computer science and biology to model and describe the interactions between intelligent rational decision-makers. Although game theory is applied to a wide range of behavioural relations, and has developed into an umbrella term for the logical side of decision science, including both humans and non-humans (e.g. computers, insects/animals), there are significant aspects that it fails to account for in human-centric technical systems, such as reputation and recommendation systems.

There are numerous theories from both the social and the computational sciences that try to tackle different phenomena exhibited in people's actions and in the social interactions among the community members in the real (traditional) and the online (virtual) world. Each of these theories makes certain assumptions when it tries to describe and model these interactions. While the number of assumptions is of less importance for the relevance of a certain theory, the impact of each assumption must be carefully studied and occasionally revised in the virtual systems to account for the evolutionary character of human and user-centric communities.

Several theories were presented in this chapter, from both the social and the computational sciences, in light of their use for the problems that are addressed in the thesis. The significance of each of them for the particular issues and the purposes of their use are described. Special attention was paid to balance theory and subjective logic, which are the major theories employed and analysed throughout the thesis. Their applicability is elaborated together with the formal apparatus for the exploration of social interactions. Balance theory is fairly simple to formalize and understand, whereas subjective logic is more empowering with its mathematical calculus for more technical domains.

Social network analysis forms the basis of any descriptive and analytical approach that

tries to tackle issues related to social interactions in virtual environments, whereas systems theory provides a sound basis for understanding, designing and maintaining any type of system of either a social or technical character. Finally, prospect theory and choice architecture can help in explaining the choice of an option and how choice decisions are evaluated when the options available do not differ significantly in their expected value or where optimal choice is not evident. They address those cases where decisions require choices among options that promote conflicting values. By examining how these options are compared and evaluated in terms of gains and losses in each issue area and not only in terms of absolute outcome, it is possible to shed new light on the critical and previously unrecognized aspects of decision-making science and practice.

3 The Upgraded Taxonomy of Trust Systems

3.1 Introduction

This chapter pinpoints the lack of a systemic perspective in the design and analysis of computational trust systems and identifies several points where trust research can benefit greatly from a holistic approach. It aims to tackle the first hypothesis as part of the research contributions of this thesis. More precisely, it is demonstrated that the current frameworks for categorization of trust systems, their subcomponents and their design principles lack important elements from a systemic perspective, thereby presenting the need for a novel taxonomy of trust systems. Taxonomy is the practice of classification and categorization of interrelated concepts and principles (Grossi, Dignum and Meyer, 2005). General systems theory is presented as the groundwork for systemic thinking, whereas case study research is elaborated as the systemic methodology used for establishing the framework as a novel perspective of building categorization theories in the field of trust systems.

In essence, guidelines for systemic reasoning about trust are provided in this chapter. With that, awareness is raised about the fact that the trust system as a whole should not be analysed only through the behaviour of its parts: considering the interconnections between those parts is also crucial. A novel taxonomy for organizing the design principles for trust systems is introduced, guided by the principles of general systems theory. The taxonomy is presented as a hierarchy of disassembled subcomponents of trust systems and categorizes the systemic properties of those systems. Thus, it provides a multidimensional map of the interdependencies among the system entities, which are proven to be lacking in the current technical solutions. It incorporates social factors into the design guidelines of trust systems by granulating the trust system components in view of four identified influencing factors.

Finally, the taxonomy is evaluated and its significance as a novel systemic perspective of the field of trust systems is established by following the principles of case study research and semantic analysis (Eisenhardt, 1989; Fortuna, Grobelnik and Mladenic, 2007). Two representative case studies of practically implemented real-world trust systems are chosen to be evaluated in relation to the taxonomy (in a non-random manner). Possibilities for improving their design and functionality are suggested based on the novel categorization of the system entities. This work is the first to apply fundamental social science principles from general systems theory and case study research for the purpose of theory building and the evaluation of technical human-centric solutions (Ažderska & Jerman-Blažič, 2012). By doing this, it joins the creative vigour of soft systemic analysis to the quantifiable practicality of "hard" technical solutions.

The chapter is organized as follows. In the next section, the progress towards formalization and systematization of the design of trust systems is reviewed, along with their separate components, and the related work is presented. Case study research is put forward as a way of defining new theories of an interdisciplinary character. The subsequent sections outline the methodology and introduce the proposed framework. Practical observations and insights from other approaches are elaborated throughout the entire chapter, enabling better granularity of the presentation and a revision of the existing

research map of trust systems. Furthermore, a semantic study is presented of the research field of trust systems. The results from this study will further lead to the selection of two cases that will then be evaluated in relation to the newly introduced framework. The two specific cases are BarterCast – a distributed reputation mechanism, and Yahoo! Answers' reputation system, which is integrated into a centralized social platform. The final section concludes the chapter with a constructive discussion, outlining future directions.

3.2 Basic Concepts and Related Work

3.2.1 Systems Thinking: General Systems Theory (GST)

The common practices of solving systemic problems vary across the areas where the problems appear and among the experts that search for their solutions: technicians strive for things to work in a desirable manner, whereas social scientists look for things to behave as desired. However, the Internet's socio-technological scene, where the interactions among machines and humans are blurred, does not allow for strict separation between those practices anymore. Any approach aimed at solving a systemic problem comprised of both human and technical entities when these are in interaction has to account for both the behavioural and the performance characteristics of those entities. Such a requirement can only be met by an interdisciplinary and holistic approach, by bringing together fragmentary research results in a comprehensive view on technology, nature and society. As a theory of wholeness, GST is the natural choice of approach for exploring trust issues and determining the desired properties of trust systems. The purpose of GST is postulation, formulation and derivation of universal principles that are valid for the systems in general. It is a theory that cuts across most other disciplines, linking closely the generalized concept of *organization* to that of *information* and *communication*. GST's founder, Ludwig Von Bertalanffy (1969), recognized in the various sciences the general tendency towards systemic integration, and the need for a general theory of organization. Thus, GST's goals are the formulation of generalized theories of system dynamics, goal-oriented behaviour and a hierarchical structure.

With the aim of defining a trust system that will work on either a fully technical level (embedded in the communication protocols between devices) or a higher level (emerging as a result of the system entities' behaviour, interfacing with the social and technical environment), GST provides a way of thinking in terms of system entities *in mutual interaction*. The developed framework specifies the system entities and the interconnections between them, providing analysis across the factors determined to influence the trust systems' design. Therefore, the position of trust systems in the general systems space is determined. This enables the developed taxonomy to resemble a systemic approach in designing trust systems[2] and facilitates the understanding of the internal relations between the RM subsystems and their connection to the external world. Moreover, it allows a common ground to be set for the widespread but scattered research on computational trust and reputation, while placing people (i.e. the users) at the crossroad of design factors and system entities.

The research on computational trust offers a myriad of formal models and quantitative analysis, without any conceptual framework of common design principles or qualitative analysis supported by multiple and diverse perspectives of an interdisciplinary nature. Thus, the human factor, as one that influences and even controls the flow in many of the system processes, has been greatly underexplored. Despite the significant body of work

[2] Although we acknowledge trust and reputation as two different concepts, we use "trust systems" and "reputation mechanisms" interchangeably. The justification for that is the general nature of these phrases, and the convenience in providing adequate expositional diversity throughout the text.

on trust and related concepts found in the social science literature, the transition from a word-of-mouth to a virtual world is far from being straightforward when it comes to applying trust reasoning. Although systemic thinking is often circumvented in the development of conceptual models, it is through such an approach that simple and efficient working principles for any system can be established, regardless of their nature (social, technical, natural). One of the few works that accounts for the interconnection between system entities through the so-called network effects is that of Akoglu, Chandy and Faloutsos (2013). The authors propose a scalable (linear with network size) algorithm for opinion fraud detection in online reviews called "FraudEagle", which operates in a completely unsupervised manner. However, the model mainly concentrates on objective criteria for scoring and fraud detection, rather than on the subjectivity of opinions and the organismic nature of the online trust systems. Furthermore, it is not concerned with the bias in the users' opinions, but it is however a candidate for bias prevention by addressing the problem of fraud detection.

The work presented in the next section aims to bring systemic thinking into the focus of designing and operating with human-centric systems. Moreover, it strives to highlight the fact that it is through systematic analysis that we can understand the difference between inherent and emergent system properties, in order to choose the meaningful action in a certain system, in addition to its design. While trust itself is one such emergent property, it is curious to observe that trust has mainly been considered a hard-wired and inherent systemic feature by the existing formalization efforts and technical implementations.

Several taxonomies have been developed in the field (Marti and Garciamolina, 2006; Huynh, 2006; Alani et al., 2004; Josang et al., 2007). However, as a categorization of system entities and relations, a taxonomy is hardly useful if it only identifies the trust system's components. Crucial in the analysis and design of trust systems are the connections between their components. Marti and Garciamolina (2006) approach trust systemically, decomposing trust systems into subsystems. However, they determine no interdependencies between the entities, processes and design factors, and consider users external to the system hierarchy. In addition to these conceptual models, there are also surveys of trust-scoring algorithms (Josang et al., 2007; Noorian and Ulieru, 2010). Such algorithms focus on one of the system components, thereby providing an incomplete view on the system functions and its internal dependencies. The presented work identifies the missing parts in the systemic research on trust and determines their application for providing trust sustainability, placing users in the core of the trust system.

3.2.2 Case Study Research

Despite the relatively widespread use of case studies as a proof of concept in the technical literature, the possibility and significance of theory building and evaluation from case studies is largely disregarded. This is very likely due to the fact that technical and engineering methods strive to base their theories on quantifiable approaches and statistically verified results, or other formalisms rooted in axiomatic and hypothetical-deductive reasoning. This, however, often prevents researchers from getting comfortable with the inductive methods of scientific inquiry, which require qualitative iterations between the analysed data/cases and the hypothesis that is being tested or evaluated (Eisenhardt, 1989). In other words, it leads to neglect of some of the core values of systems thinking comprised of iterating between the ideas one has and the experiences that come from applying those ideas – a process resembled by the soft systems methodology (SSM) (Checkland, 1981; Checkland and Scholes, 1999). As a research strategy, case study research follows closely the soft systems methodology.

According to Yin (1981, 2): "*The distinguishing characteristic of the case study is that it attempts to examine a contemporary phenomenon in its real-life context, especially when the boundaries between the phenomenon and context are not clearly evident*". Hence, a case study differs from quantitative approaches, which essentially separate the phenomenon from its context. However, the upside of the process of establishing and evaluating theories from case studies (i.e. the opportunity to open new avenues of thought) is also the trait that pushes the borders of scientific uncertainty too far for some researchers to get along with it. This was carefully analysed and discussed by Flyvbjerg (2006), who tried to shed new light on five misunderstandings about the nature and value of case study research. His work resulted in the following revised list that translates the five misunderstandings into a set of useful insights about case study research:

1. *Predictive theories and universals cannot be found in the study of human affairs. Concrete, context-dependent knowledge is, therefore, more valuable than the vain search for predictive theories and universals;*
2. *One can often generalize on the basis of a single case, and the case study may be central to scientific development via generalization as a supplement or alternative to other methods. But formal generalization is overvalued as a source of scientific development, whereas "the force of example" is underestimated;*
3. *The case study is useful for both generating and testing hypotheses but is not limited to these research activities alone;*
4. *The case study contains no greater bias toward verification of the researcher's preconceived notions than other methods of inquiry. On the contrary, experience indicates that the case study contains a greater bias toward falsification of preconceived notions than toward verification;*
5. *It is correct that summarizing case studies is often difficult, especially as concerns case process. It is less correct as regards case outcomes. The problems in summarizing case studies, however, are due more often to the properties of the reality studied than to the case study as a research method. Often it is not desirable to summarize and generalize case studies. Good studies should be read as narratives in their entirety.*

The provided list not only justifies the reliance on case study research as part of the methodology, but also points to the way the study should be understood. By merging systemic thinking with case study research, a solid ground is provided for performing a case study research that testifies to the validity of the five revised misunderstandings defined by Flyvbjerg.

3.3 Identifying Missing Elements in the Trust Taxonomy through Jordan's (GST) Taxonomy

Among the most prominent works on system categorization that apply the principles of systems thinking is the general systems taxonomy (Jordan, 2003).

Table 1: *Organizing principles of Jordan's taxonomy.* The categories to which trust systems are assigned are italicized.

Rate of change	**Purpose**	**Connectivity**
Structural (static)	Purposive (system-directed)	Mechanistic (non-densely connected)
Functional (dynamic)	*Non-purposive (environment-directed)*	*Organismic (densely connected)*

Jordan's taxonomy has three organizing principles: *rate of change*, *purpose* and *connectivity*. Each principle defines two antitheses, resulting in the three pairs of properties shown in Table 1. Trust systems are assigned to the categories that are italicized. Justification of this choice follows.

3.3.1 Rate of Change: Dynamicity (D)

Static systems are those that exhibit no change in a defined time span. It is reasonable to argue that any system can be seen as both static and dynamic, depending on the time interval within which the changes are observed. However, trust systems are expected to provide long-term incentives and to support users in decision-making in a dynamic manner. To do that, they consider the quality of past experiences and the history of entities' transactions, but provide decision support under the present dynamics. Thus, trust systems are directly concerned with the need for good information propagation and result convergence (for proper signalling), as well as timely response and prompt fault/error detection (for proper sanctioning). Furthermore, the trust system entities are interdependent or temporally interact through interlocking dependencies, leading to increased complexity of the system, directly affecting its stability and robustness. On the one hand, these properties reflect the extreme dynamicity of trust systems. On the other hand, they also reveal the need for an evolutionary approach in the system design and analysis in order to cater for the high rate of change of the entities' behaviours in relation to the context in which the system itself operates. Although *dynamicity* as described here is considered a system property, it is addressed more broadly later after the introduction of *time and dynamics* as a design factor. The rate of change plays an important role in understanding and describing the trends of unexpected users' behaviours, such as: herding, information cascades, conformity, mass hysteria, etc. (Burt, 2000; Leskovec et al., 2007; Watts, 2002). Special attention is paid to these issues in the formal analysis of biased user behaviour in the following chapters of the thesis.

3.3.2 Purpose: Environmental Orientation (E)

The principle of *purpose* determines the flow direction and exchange of energy/information within the system and with the environment. The two possibilities are a *system-directed* flow and an *environment-oriented* flow. The former tends to maintain stable conditions within the system, whereas the latter modifies the system to obtain a desired state or bypass certain disturbances. Although human-made systems are usually purposive, human-centric systems whose state and output depend essentially on social factors tend not to be. Makridakis and Taleb (2009) have addressed the issue of limited predictability and the high level of uncertainty in various areas of science, and in life in general. There is a remarkable body of empirical evidence speaking about the disastrous consequences of inaccurate forecasts, which are distinguished by three types of predictions: those relying on patterns for forecasting, those utilizing relationships as their basis and those for which human judgment is the main determinant of the forecast. Trust systems are designed to rely on all three. They are human-centric and utility-driven, and are expected to perform under conditions of high risk and uncertainty, which categorizes them as *environment-directed*, i.e. *non-purposive*. Clearly, categorizing systems according to the principle of purpose depends on the "standpoint" of the observer. One of the goals of the thesis is to examine the possibility of distributing the control and the learning processes throughout the system by considering the human factor as the driving force. This goal links the first with the fourth component of the hypothesis outlined in the Introduction; namely, the need for a novel taxonomy of trust and the role of user bias in

the process of information aggregation. To do that, the boundaries of the trust system are opened, and the binding role in keeping the system and the environment as a single organizational whole is ascribed to the users. In other words, trust systems are not seen as closed and residing within a single context, but as ones whose goal changes dynamically in relation to the entities' behaviour and the external environment. Therefore, the aim of distributing the control among the system entities comes as a need to provide adaptability in a system whose state is determined by self-organization, rather than by predetermined goal-seeking behaviour.

3.3.3 Connectivity: Dense Connectivity (C)

In *mechanistic* systems, the removal of parts or connections does not affect the remaining components. In contrast, *organismic* (densely connected) systems change even when a single connection changes. Trust systems are complex and depend heavily on the interactions among their entities. They are inherently of a non-linear nature, and the outcome of each interaction has no predictable impact on the overall system. This does not imply that a linear method cannot produce acceptable results. Although many proposed models rely on the ability to determine a general reputation score for an entity, there has been no evaluation in the literature of the impact of error-prone trust systems on the general system welfare. This comes as no surprise, as the non-linear, complex and *densely connected* nature often allows only for retrospective analyses of the result, rather than building a successful predictive model. It is worth noting here that, for reputation mechanisms deployed on a strictly technical level (for example, for routing optimization, resource allocation and workload distribution), the systematization with respect to the *connectivity* property would be different; such systems would be considered *mechanistic*. The reason for this is that most of the technical parts of the systems have predetermined ranges of operation and protocols for information/data exchange at the given system level. This is not a claim that the physical processes are of a deterministic nature, but that they are more lendable to predictive analysis for determining the outcome and behaviour of the systems' parts and processes.

Luhmann (1979) recognized trust long ago as a means of reducing complexity in traditional social systems. However, it is not trivial to generalize this as an implication for the technical solutions, which are much more scalable, in addition to being highly dynamic. Although most of the proposed trust models provide some discussion or evaluation of their scalability (among other performance criteria), dynamicity and scalability are mainly considered technical nuances of a given solution and are analysed separately from the interactions between the system entities. The life span of the natural (organismic) systems spreads through two major and subsequent phases: growth and maintenance (Odum and Barrett, 2004). The resources that systems use in the first phase are mainly intended for growth; in the second phase, on the other hand, the systems' resources are employed for maintaining the state and preserving/improving the quality of operation (a quick recall of the human as a system would suffice to realize this). This transition from a quantity- to a quality-based operation is often seamless, prospectively unpredictable and only retrospectively realizable, which is one of the things that nourish the ever-present mechanistic trend of resolving systemic issues.

The significance of GST is the clarification and simplification of the often-misused concept of *a system*. By using sufficient generality and simplicity in this thesis, trust systems are established and categorized as *dynamic (D)*, *environment-oriented (E)* and *densely connected (C)*.

3.3.4 The Proposed Taxonomy of Trust Systems: Redefining the Impact Factors

This section moves towards disintegration of trust systems into their building components, identifying the interdependencies between them and determining the influencing factors that guide the system design. The results can be summarized as follows:

- It categorizes common and important concepts of trust systems, thereby establishing a common systemic vocabulary;
- It provides a novel and holistic approach in the multidimensional mapping and assessment of the completeness and consistency of trust systems;
- It introduces the missing granularity in the current taxonomic map of trust systems;
- It employs the identified D-E-C nature of trust systems described in the previous section to detect the factors that influence the system design, thereby making the proposed framework complete.

Table 2: *Breakdown of the trust system components (Marti and Garciamolina, 2006).*

Trust Systems		
Information Gathering	**Scoring and Ranking**	**Response**
Identity Scheme	Good vs. Bad Behaviour	Incentives
Information Sources	Quantity vs. Quality	Punishments
Information	Time dependence	
Aggregation	Selection Threshold	
Stranger Policy	Peer Selection	

As a skeleton for initial decomposition of the system, Marti's taxonomy as presented in Marti and Garciamolina (2006) is used, with the categorization of the trust system components given in Table 2.

In addition to the subsystems and factors outlined in Marti's taxonomy, the resulting framework (presented in Figure 3 – Figure 6) allows direct mapping of the models across the factors dimension and the subsystems dimension in a consistent manner. Presented through a series of taxonomy trees, it captures the subsystem's granulation to a greater extent, demonstrating the attributed multidimensionality of the framework. Hence, by determining the depth to which a certain property does or does not satisfy the requirements presented by each of the factors, it is possible to immediately assess the impact that the analysed property has on the overall system. For example, a missing property in the leaves has a lower impact than a missing property in the inner nodes of the taxonomy tree.

As the *system architecture* has already been analysed as a design factor in Marti and Garciamolina (2006, 7), it is now complemented by the four new factors to fully capture the D-E-C nature of trust systems. The D-E-C properties take into consideration the non-linearity introduced by and met in the trust systems and the emergent properties of the entities' interactions. This section redefines the factors that are the constructs of the framework.

3.3.4.1 Active Entity Behaviour (AEB)

Active entity behaviour (AEB), as presented in Figure 3, deals with the impact that the entities' behavioural changes have on the information and energy flow. As a design

factor, it *accounts for the D and E nature* of the trust systems. Although some technical aspects of reputation mechanisms attributed to AEB come as challenges in coping with the network dynamics, the importance of anticipating AEB in human-centric systems goes beyond the understanding of its technical implications. A system burdening the users with enrolment conditions and inconsistent participation requirements, or hindering their access to resources, may encumber the users to the extent of edging them out. Trust systems must exhibit high adaptability when addressing the outlined issues. Both the hard technical and the soft usability aspects are important as solutions. The former includes availability/connectivity checks for forming overlays of reliable entities, while the latter requires bootstrapping techniques for the new-coming entities, incentive policies for the experienced users, and reputational, institutional and societal "pressures" as the system scales up.

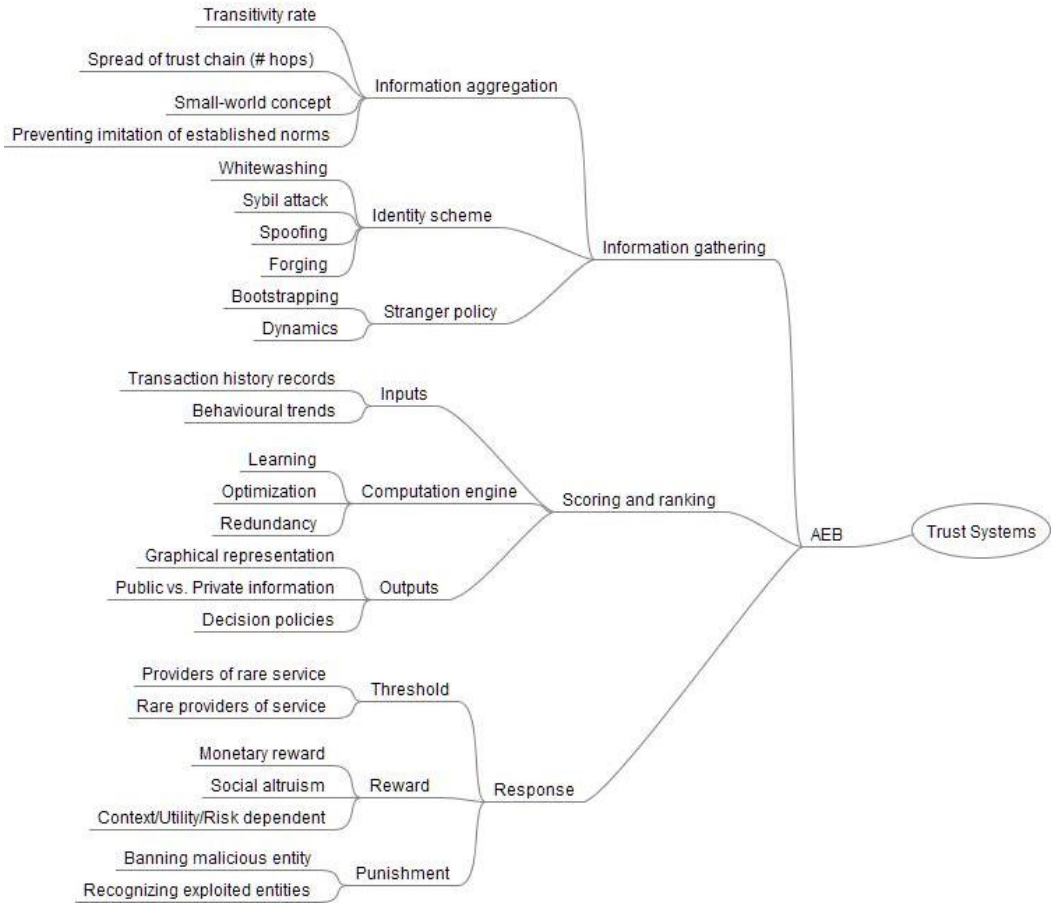


Figure 3: *The interdependencies between active entity behaviour and the trust system's subcomponents.*

Scalability is often seen only as an architectural issue, but human-centric systems do not have to be technically large scale to require systemic intervention. As users' activity shifts from individualistic to group activity, the different mechanisms are needed in order to cater for the collective actions of the community members. In Zhang (2011), the author studies the anchoring effect of item ratings on user preferences. The author shows that users' inclination to provide positive feedback is amplified when the item rating is publicly visible. By imitating group members, especially when having no determined preferences, users contribute to the so-called group polarization, bandwagon effect or herd behaviour (Kahneman, 2003). A study on group polarization on Twitter also showed that, when part of a group, people form opinions and take actions that are more extreme

than if making individual choices (Yardi and Boyd, 2010). Understanding the effects of such bias phenomena exhibited by the active entities is even more important if one considers the contradictory goals of the commercial (marketing) interests and the computer science community. While the earlier aim was to exploit the formation of such information cascades for better product diffusion and viral marketing (Lieberman and Montgomery, 1988; Aral and Walker, 2012), the later aim was to design efficient algorithms for aggregating information coming from dependent and conflicting sources (Dong, Berti-Equille and Srivastava, 2009; A. Josang et al., 2003; Smets, 2007; Zhao, Rubinstein, Gemmell and Han, 2012), and to study the role of social networks in information/product/behaviour diffusion (Bakshy et al., 2011, 2012; Goel et al., 2012; Myers, Zhu and Leskovec, 2012).

All of the above implies that the impact that the community structure and its evolution have on the operation and performance of trust systems is crucial for the design throughout the system maintenance. More specifically, it is the community structure and the behavioural trends exhibited by the active entities that dictate many of the trust properties that are mainly taken for granted and validated by the current models. These include the extent of transitivity in the trust propagation chains, the depth to which trust information influences the chain and the presence of an authority figure that controls some of the information flows in the process of information aggregation. These are also the key factors considered by the malevolent entities in their efforts to compromise the work of the trust system. Clearly, the policies for punishment and reward also have an impact on the entities' behaviour, as well as the motives for that behaviour. Finally, the representation schemes of the trust information (public or private, numerical or descriptive, binary or multi-valued) and the visual appearance of the system elements used by the users as a type of active entity can have a significant impact on the outcome. It becomes evident that not only are all of the subsystems related to one another, but the transformation of the system inputs to system outputs can take many different paths because it is processed by a certain subsystem.

3.3.4.2 Context

Reputation and trust become meaningful only if treated as a domain, i.e. context-specific. As a set of circumstances surrounding a particular situation, *context* as a design factor (Figure 4) *accounts for the D and E nature* of trust systems, and is directly related to the previous definition of AEB. Despite the various types of trust defined in the literature, only a few approaches consider the context dependence. Among the most well-known are the efforts to incorporate semantics into the reputation model (Golbeck, 2006; Heath, Motta and Petre, 2007). Most of them employ the techniques for content filtering, and none considers the impact of context on the separate trust system components. For instance, the locality of trust information needed to form an opinion from the social networks of friends and acquaintances largely differs from the one needed for public e-commerce or recommendation systems. It often suffices to consult one or a couple of friends (directly) before forming an opinion about the trustworthiness of a certain entity, as opposed to the number of (indirect) recommendations that would provide a similar level of certainty for one's inquiry in that context. By making use of contextual information in the reputation evaluation, not only is the level of the entities' expertise obtained, but also the domain where this expertise is relevant. To address this issue, implementing methods for assessing context similarity with respect to a given criterion is a viable solution. Such methods enable the introduction of a weighting scheme for the reputation relevance according to the shared contextual parameters between the entities. The result produced in this manner could be used both as a measure of the impact in a reputation assessment and as an interoperability enabler between different systems (which

are not necessarily trust systems).

There is a wealth of concepts intimately related to trust, such as distrust, disbelief, suspicion, altruism, rivalry, etc. While trust is relatively non-transferable across dissimilar contexts, it is worth examining whether the same stands for the other concepts. There are strong indications from the real world that distrust tends to be more transferable across contexts. Moreover, it also exhibits different propagation ability through the relationship chains between entities (Victor, Cornelis, De Cock and Herrera-Viedma, 2011). In order to capture the innate subjectivity of contextual trust reasoning, the apparatus employed to infer an entity's trustworthiness has to be embedded in diverse formal models that are able to work with different representations of the trust information: numerical, visual, linguistic or otherwise encoded. An insignificant amount of research has investigated these issues, which is one of the reasons we enlist our work into the research on *signed networks*.

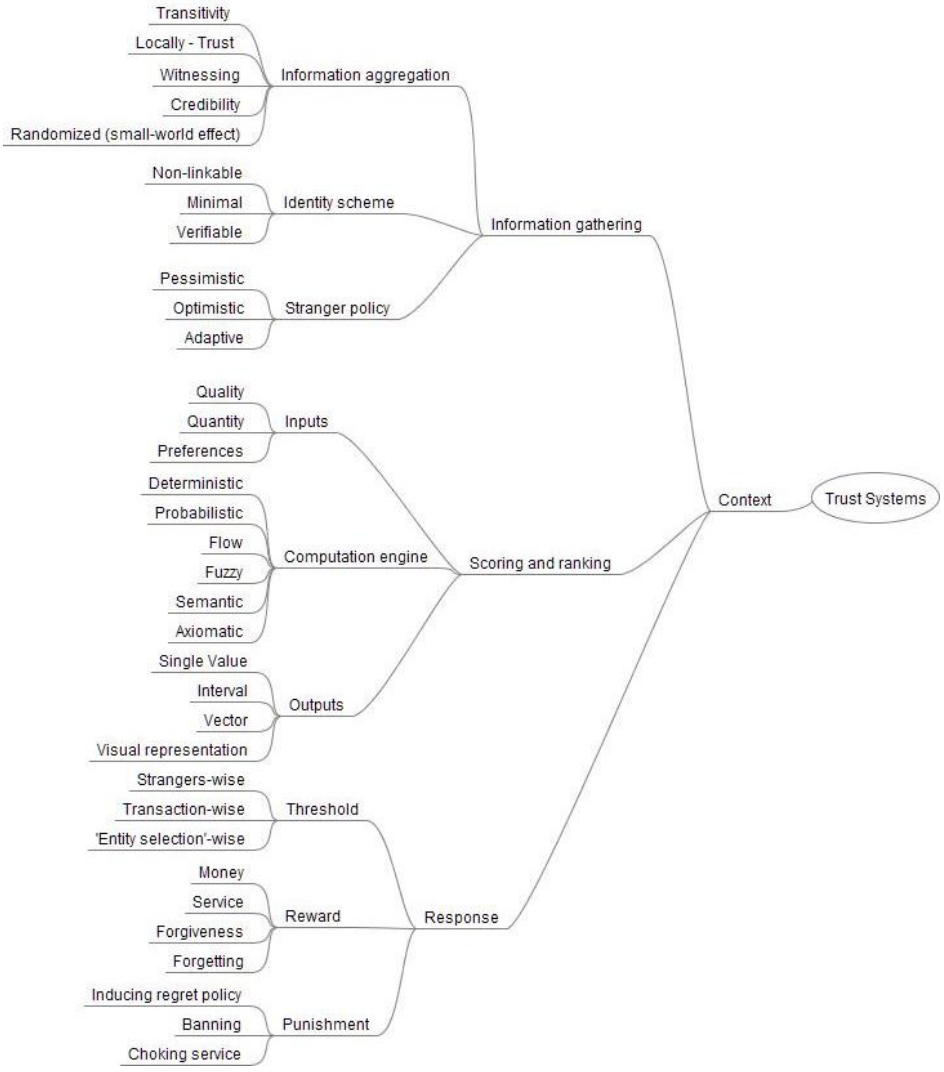


Figure 4: *The interdependencies between context and the trust system's subcomponents*.

Clearly, a designer could easily be lost in this "labyrinth of system requirements". But this is also the point where the introduced notion of "trust as an emergent property" can be appreciated; this new emergent property implies that trust is created in accordance with the context within which an interaction takes place. In that context, it is neither useful nor reasonable to claim predictability of the outcome from a trusting behaviour, or moreover,

to hard-wire a model of trusting interactions. This is especially important to keep in mind when talking about the robustness of a trust system.

3.3.4.3 Resilience and Evolutionism

"Uncertainty paralyses effort [...] One of the chief signs of a high civilization is the reduction of risks and the lessening of the many perils of life and property to which human beings are exposed" (Fisher, 2011).

The decision of which contextual elements need to be selected as part of a system should be based on analysis that connects the trust context to risk and uncertainty. Risk is conceived as the possibility of triggering unexpected, unlikely and detrimental consequences by means of a decision attributable to a decision-maker (Luhmann, 2005). The extent of uncertainty, expected utility and cost of performing an action influence the perceived risk. The circular, interlocking and time-dependent relationships among the trust system components give rise to properties manifested in the general system, but that may not always be the properties of the components. Thus, the behaviour of the system cannot be explained in terms of the behaviour of its parts. This is a common property of complex systems that depend on social dynamics. *Resilience and evolutionism*, illustrated in Figure 5, consider these properties, *accounting for the D, E and C nature of the trust system*.

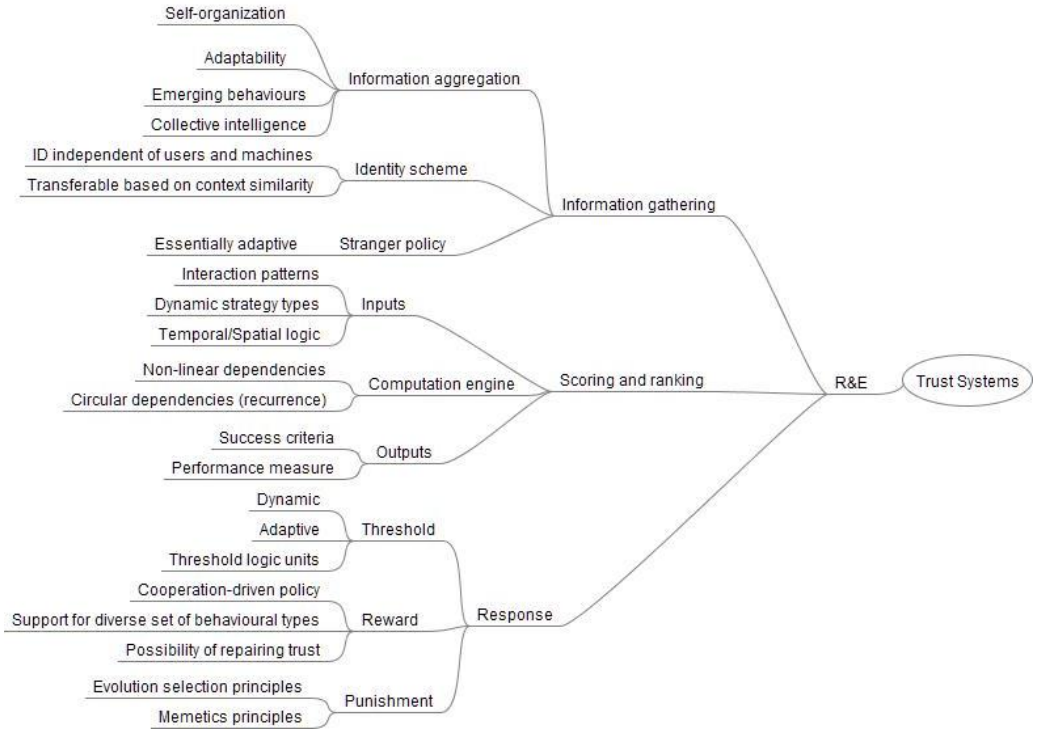


Figure 5: *The interdependencies between resilience and evolutionism and the trust system's subcomponents*.

Many of the known approaches that offer a formal description of trust in a technical environment use the MAS (multi-agent systems) model, which is based on the assumption of the existence of multiple interacting agents as active entities. It is assumed that the control is imposed from outside, but the intelligence in the system is ascribed to the comprising components. Considering the D-E-C nature of RMs, the MAS model is restricted in the sense that it does not consider issues like insufficient adaptive capacity (in terms of communication, cooperation, spatial and temporal organization), self-organization and complexity. The system as a whole is expected to have high resilience in

cases where some components are exposed to malicious exploitation, or when a system entity exhibits a sudden behavioural change. Furthermore, MAS models exist mainly as a concept for modelling and simulation, rather than a practical implementation. A class of systems satisfying the aforementioned requirements is known as complex adaptive systems (CAS). The idea of a CAS approach in the trust system design has not been investigated in any depth so far, despite the growing literature. It is an evolutionary approach for tackling resilience and dynamics issues in complex, socio-technical environments. It treats intelligence as an emerging property from the entities' interactions, accounting for the learning loops, the distribution of control and responsibility, and the self-organization resulting from the system evolution.

3.3.4.4 Time and System Dynamics

From definitions 1 and 2 follows the assumption that trust and reputation have very different dynamics; they do not even vary proportionally. This discrepancy cannot be captured by any model, as both trust and reputation are in reality intangible matters. The dynamics embedded in the current models comes from time-discounting the relevance of the gathered information. However, such an approach disregards the importance of past information in terms of impact on the outcome. Discounting by recency and frequency is different from discounting by impact. The former gives very little credit to users who take important actions at a lower rate, which is against the incentive scheme for good behaviour. Relations between reputation and time have been studied extensively in the literature; however, *dynamics* as presented in Figure 6 is insufficiently considered as a factor influencing the design choices. Such analysis would *account for the D and C nature of trust systems*.

To better understand the difference between *time* and *dynamics*, it is useful to think in terms of *change* and *rate of change* (as their analogues). From a social perspective, a proper time policy would enable incorporation of the traditional analogues for *forgetting* into the reputation score. However, it is when the time is coupled with the dynamics that *forgiveness* and *regret* can be considered implementable properties (Marsh and Briggs, 2009).

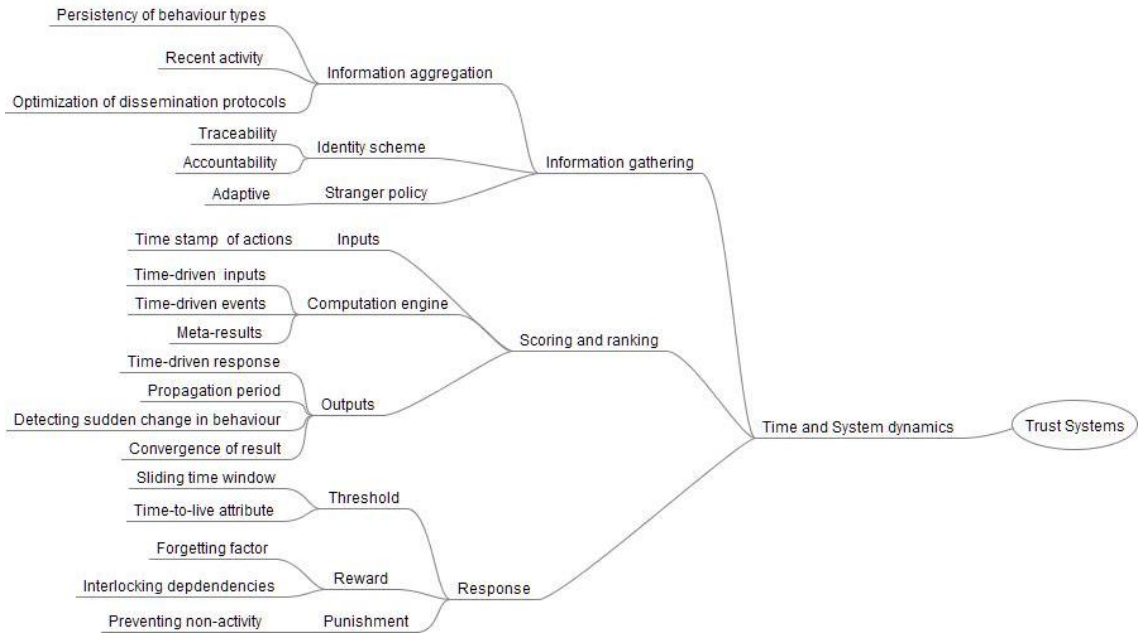


Figure 6: *The interdependencies between dynamics and the trust system's subcomponents*.

Despite information about the general trustworthiness regarding the entities' expertise,

insights into their experience in the claimed quality are also obtained if dynamicity is taken into account. Such an approach would provide a shift from information-based to knowledge-oriented design of trust systems. Resolving these issues depends on the subsystem they appear in, and involves decisions regarding ID permanency, information filtering, data retention, time-stamped actions, reputation convergence, synchronization and reputation updating. One way of addressing them is by introducing a sliding window within which reputation gains certain importance (possibly threshold-controlled), by discounting the (meta-)results obtained at certain instances, or by a combination of both. Finally, a trust system that works well for a small community (or for the initial members of the community) needs to be modified over time as the community grows and the interaction patterns change. Popular and successful systems are more likely to encounter abuse.

3.4 Case Studies through the New Framework

In the tabular matrix representation of the taxonomy given in Ažderska and Blažič (2011), the current tree roots were attributed to one of the dimensions. The subsystems defined in Marti's taxonomy represented the second dimension, whereas it was proposed that the description of the properties and their presence/absence in a trust system (represented by the leaves in the current taxonomy) were composed of the matrix cells. This kind of representation enables the discovery of important missing constructs in the analysed case studies, but it also led to the improvement of Marti's taxonomy.

- A more granular structure allows for diversification of the roles of the trust system components, but more importantly leads to more in-depth analysis;
- New components are obtained from the iterative analysis of the case study (performed on a single trust system).

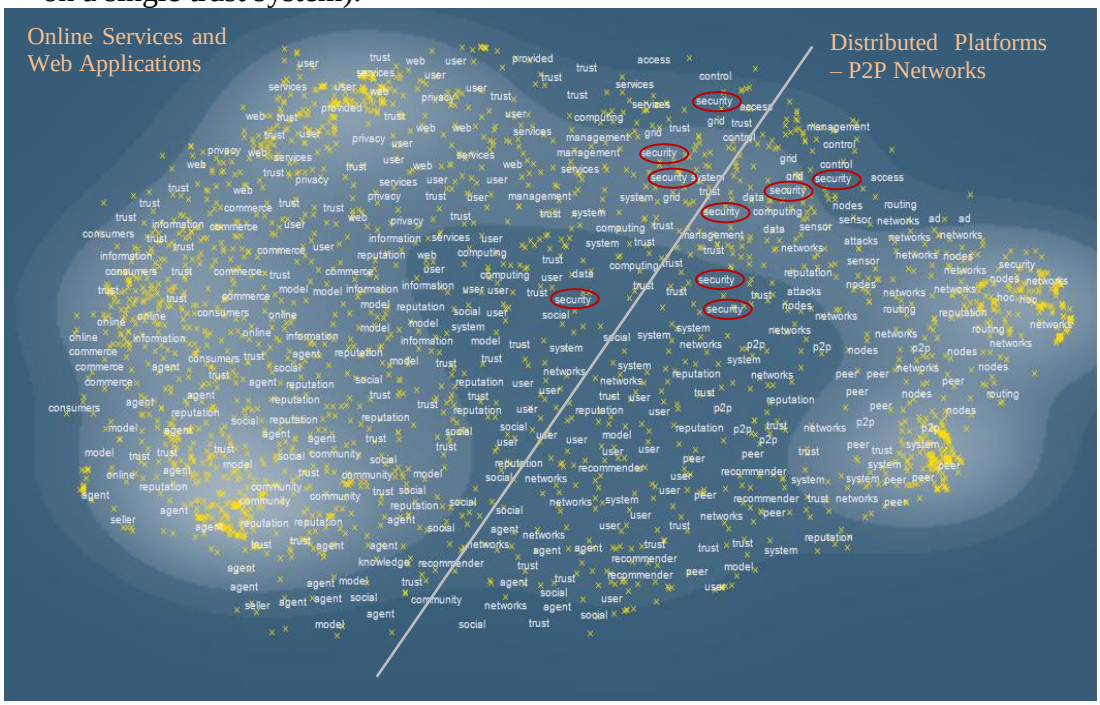


Figure 7: *Ontological map of the ACM DL query "trust reputation systems".*

As a consequence, a new construct in the form of a decision-tree appeared to be crucial as a requirement for the design of densely connected and dynamic trust systems. It enables qualitative assessment of the immediate impact of the property fluctuation on the overall system performance. This allows the taxonomy to be represented as a tree

structure. This structure allows the impact of a certain systemic property to be immediately assessed according to the depth at which it resides on a particular branch and under a particular subcomponent.

In order to select appropriate cases for application of the new taxonomy framework, in-depth semantic analysis was carried out to select the most suitable representative cases. The outcome of this analysis is an ontology of the state of the art in the area of trust and reputation, presented in Figure 7, which shows the major areas where this research is concentrated.

A semi-automatic ontology editor (Fortuna, Grobelnik and Mladenic, 2007) was used for analysing a corpus of 2405 documents, consisting of titles and abstracts of papers from ACM DL. The visualized ontological map shows two clusters grouped around *online services and web applications*, in addition to *distributed platforms and P2P networks.* The former (left) cluster consists of concepts like *agent*, *services*, *web*, *social*, *recommender*, *virtual* and *business*, whereas the latter (right) cluster is distinguished by *peer*, *networks*, *nodes*, *attacks*, *wireless*, *sensors,* and *ad hoc*. Although sometimes the title and the abstract are an incomplete representation of the document's content, a short summary of a paper's goals and results provides sufficient information for extraction of the necessary generalities. Furthermore, the following outcomes from the visualization can be observed:

a) The two clusters are not separated strictly from one another, despite their different architectural types (centralized and distributed). A binding concept between them is *security*. This is an intuitive support to the fact that the basic type of trust that must be present before any other is the *trust in the underlying infrastructure (system trust)* (Pennington, Wilcox and Grover, 2003), which is provided by security mechanisms;

b) The blurred separation between the two clusters reveals the possibility that research in either of the fields is mainly done separately, without mixing the technicalities of security services with the soft methods that make trust more technically graspable;

c) After querying the document corpora for "control system" and "system theory", only three documents were retrieved that had used *control* in a context other than *access* control. However, only one of the three documents used *control* in a systemic sense, but always in a context different from trust and reputation. Hence, no document was found that takes a systemic or any other type of approach compliant with general systems theory.

The findings of this semantic analysis pointed to the most suitable types of cases to be analysed to demonstrate the applicability of the designed framework: the first case belongs to the area of distributed networks, and the second to the area of online services and web applications. The specific choices for trust systems are BarterCast (Meulpolder, Pouwelse, Epema and Sips, 2009a), a fully distributed trust system in the BitTorrent client Tribler (Pouwelse et al., 2008), and Yahoo! Answers, a social platform for knowledge sharing with a centralized RM (Zaltzman and Le, 2007). These cases were chosen because:

- They are both working systems, already integrated into the hierarchy of bigger platforms;
- Their main in-built properties are social phenomena, such as friendship, trust and sense of community;
- They are directly faced with the requirements for coping with network and user dynamics; providing timely service and response; exhibiting context awareness of the actions of their entities; adapting to the ever-changing environment and requirements that come from active entities; accounting for the emergent properties resulting from

the interactions between the entities; and moreover, preserving the privacy and/or anonymity of users, to the extent of separating their actions from their identities;
- For both trust systems there are valuable sources of documentation about their design and post-implementation experiences. Tribler is in fact an open-source platform and (at the time of writing) its source code is publicly available under the LGPL Open Source licence[3]; and
- They are easily accessible and allow a "neat interaction" between our expert knowledge (for observational purposes) and their performance and work.

3.4.1 BarterCast

The design of BarterCast and the way it maps onto the framework is evaluated through the influencing factors for each of its subsystems. This is presented in the text that follows.

BarterCast (henceforth denoted as BC) is a distributed RM deployed in the BitTorrent client Tribler (Pouwelse et al., 2008). BC's goal is to combat free-riding, a problem manifested as a disbalance between the contribution and consumption of system resources.

Information Gathering. BC introduces permanent secure peer identifiers (PermIDs), based on a public key scheme using elliptic curve cryptography. To prevent spoofing, a challenge-response mechanism is implemented for validating the PermIDs. The social network creation in the Tribler client is also facilitated by the ability to import contacts from other social networks in which the peer is a member, like MSN or Gmail. Every user is referred to by a pseudonym. To provide users with some possibility of preserving privacy, the system supports the "erase from profile" option. Long-term context information is stored in databases (MegaCaches), enabling the existence of trust-based social groups. To conduct content discovery and recommendation, the system maintains a semantic overlay consisting of peers with similar tastes, called "taste buddies". These are discovered by an epidemic gossiping protocol. The connection to other peers is realized through two processes: 1) exploitation, where a peer exchanges social networks and preference lists with one of its buddies, or 2) exploration, where information is exchanged with a new, randomly chosen peer. In BC, direct experience is represented by the history of transactions (the aggregated amount of service) that peers have exchanged during the process of gossiping about their contributions (donated upload bandwidth). Furthermore, peers maintain both a private and shared history of transactions and subjectively calculate from it a reputation score when they choose a transaction partner. The private history reputation is based on entities' mutual interactions, whereas the shared history considers the interactions among all entities. To limit the computational overhead, BC considers paths of at most two hops. This simplification results from the fact that in file-sharing networks, 98% of peer pairs exchange data either directly or with a common third party, demonstrating the presence of the so-called small-world effect (Newman, 2000).

Scoring and Ranking. The network of interacting entities in BC is represented as a graph, with the peers as nodes and directed edges that represent the aggregated amount of service. Peers exchange information with others only about their direct experience. The total number of bytes transferred from one peer to another in the past represents a measure of the performed service. As input statistics, both the quantity (upload in MB) and the quality (the positive contribution) of the service are considered in the scoring algorithm. The private and the shared history forms the peer's local graph, which is used as an input to the maxflow algorithm. It computes the maximum flow over all possible

[3] http://www.tribler.org/trac/wiki/TriblerForDevelopers

paths, from a source node to a sink (target) node. The result is the highest value of (subjective) reputation that the source can give to the sink without violating "reputation capacity" constraints.

Response. BC introduces a few different types of incentives during the reputation process. First, to make use of the social groups formed in the community, it implements a cooperative downloading protocol that allows members who trust each other to cooperate and improve download performance. Second, in addition to the BitTorrent's tit-for-tat (which gives peers a short-term incentive to upload, but not to continue sharing the file afterwards), BC incorporates long-term incentives by implementing a ranking policy, which allows interested peers an initial cooperation in the order of their reputation. Third, it cherishes the peers' sense of community, which in the long run acts as a social norm for contributive behaviour. Finally, by introducing costly operations in the procedures to use system resources, BC edges out malicious peers, giving an additional incentive to regular peers to continue contributing. In order to select interacting partners, BC introduces a banning policy. The choice of whom to allow to use the resources is made according to the peers' reputation, where a reputation is required to be above a negative threshold. The reason for a negative value is to differentiate between strangers with a clean record and peers with a bad reputation.

Stranger Policy. Strangers are tackled mainly by the bootstrapping process in the Tribler client in two ways. First, to obtain an initial list of neighbours, immediately after joining, peers use a set of pre-known super-peers to bootstrap into the network. Second, there is an overlay swarm with no central component that can also be used for initial bootstrapping, content discovery and other information exchange.

3.4.1.1 Proof of concept

BC's mapping to the framework is presented in Figure 8 – Figure 11. The results suggest a space for substantial improvements in a wider set of design requirements.

Active Entity Behaviour (Figure 8). A very important aspect that BC fails to account for across all the factors in the information gathering subsystem is performing an integrity check, opinion weighting or any type of weighting of the reputation entities, their actions and relationships. In order to address this problem, BC may introduce a witnessing policy component, similar to the one proposed by Kamvar et al., (2003). This will significantly improve the credibility of reputation information obtained by its entities, but also of the final reputation score based on the gathered statistics.

Although BC combats free-riding and prevents ID spoofing, its open and distributed nature makes it a desirable target of Sybil attackers. This type of attacker is able to control multiple identities in order to consume a great amount of resources from the system. At the point where this attack becomes impossible to deal with, to the extent of exhausting system resources, it can even turn into a distributed denial of service. Moreover, because of the relatively low cost of entering the system, the ability to perform whitewashing (re-entering a system under a new identity) further lowers the credibility of the obtained reputation score, and also de-incentivizes the users from putting effort into building a long-term reputation record. In order to deal with the described issues, it is necessary to impose some kind of cost in different parts of the reputation mechanism. For instance, a (re)CAPTCHA before downloading a certain file, or solving a puzzle before using a service (Bocan and Fagadar-Cosma, 2005), will not only make the system more resilient and secure, but it will combat malice in a computationally efficient way, while also incentivizing contributing behaviour.

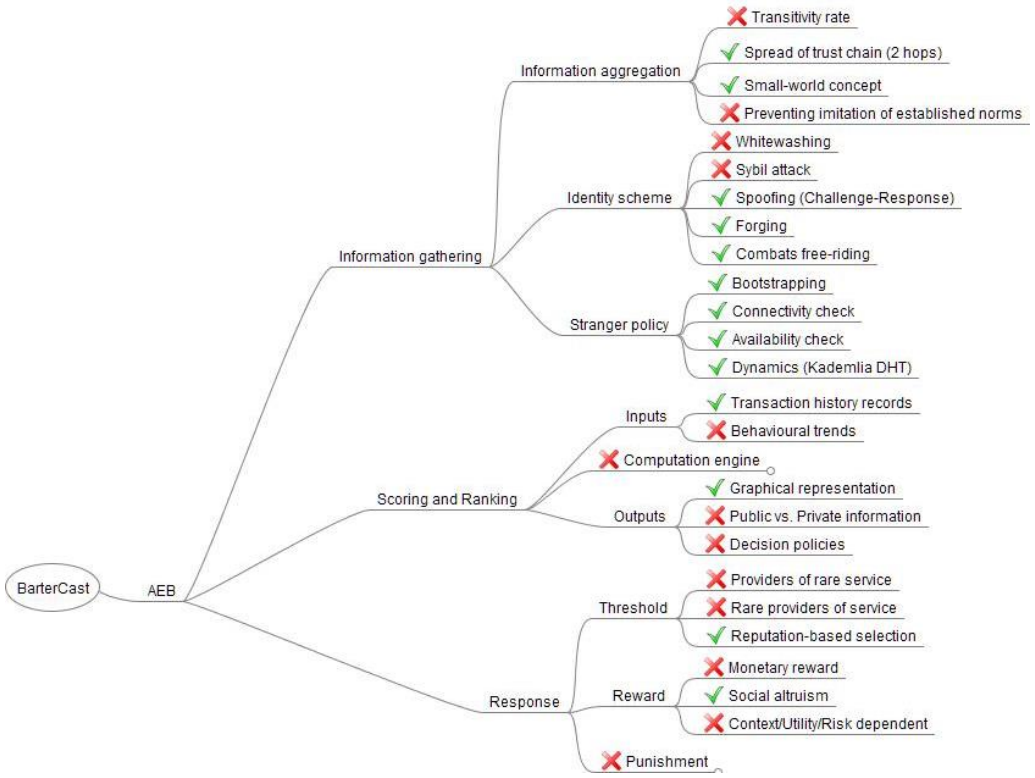


Figure 8: *Evaluation of the interdependencies between active entity behaviour and BarterCast's subcomponents.*

Context (Figure 9). Despite the fact that BC uses the idea of *taste similarity*, *taste* is much more subtle than preference. Capturing the dependence between the taste of the system entities and the context within which an action was performed or an opinion was given requires metrics for context similarity and context granularity/generality.

The purpose of such a metric would be to represent the relevance of opinions about the entities' trustworthiness, and, hence, the importance that the reputation scores calculated from these opinions will have in the final decision for taking action in the given context. Furthermore, it is crucial to consider the impact of the entities' efforts to employ certain means for good impression management in a system that relies on the aggregate effort of its members. Not doing so leads the incentive-incompatible trust systems to be easily exploitable by self-interested individuals. Designed with the aim of preventing free-riding, incentive compatibility is the paramount feature that BC should ensure. Its current design relies only on the properties embedded in the underlying BitTorrent (like optimistic unchoking and fair sharing ratio), but these are more mechanism-oriented rather than socially inspired. To strengthen the social element of content sharing, BC should rely on a more distributed responsibility system (similar to Ba's (2001)) in addition to the distributed file exchange protocol. One huge benefit from such a system is that impersonal online transactions (as the great majority in BC are) can be supported while preserving entities' anonymity.



Figure 9: *Evaluation of the interdependencies between context and BarterCast's subcomponents.*

Resilience and Evolutionism (Figure 10). BC is based on the premise that communities in the real world work well with millions of users, despite not being resistant to all forms of cheating. However, this does not speak about the impact these entities can have on the overall (computational) system welfare. Only a small percentage of peers in a file-sharing community contribute the largest amount of resources in the network. False self-representation and collusion can have an impact on the cost that largely outweighs the benefit of designing and maintaining an RM. Despite exploiting the small-world phenomenon for better gossiping in BC, this phenomenon is not an indication of an organizing principle of the nodes in the network. There is a specific structure that a network should have before applying the small-world concept: there should be a power law dependence between the number of short-distance and long-distance connections (Kleinberg, 2000; Sandberg, 2005). Otherwise, node positions can be rearranged by a switching protocol, without changing the nodes' connections (Sandberg, 2005). Neither the requirement nor the network's organizational principles were tackled or discussed for BC. Its two-hop gossiping may lower the traffic overhead for performing computations over the information obtained by the message exchanging, but it does not necessarily improve the overall performance of the mechanism. As accuracy is critical for system performance, full gossiping among the nodes may be well worth investigating. It will lower the computational overhead and lead to better convergence and propagation of the reputation score.

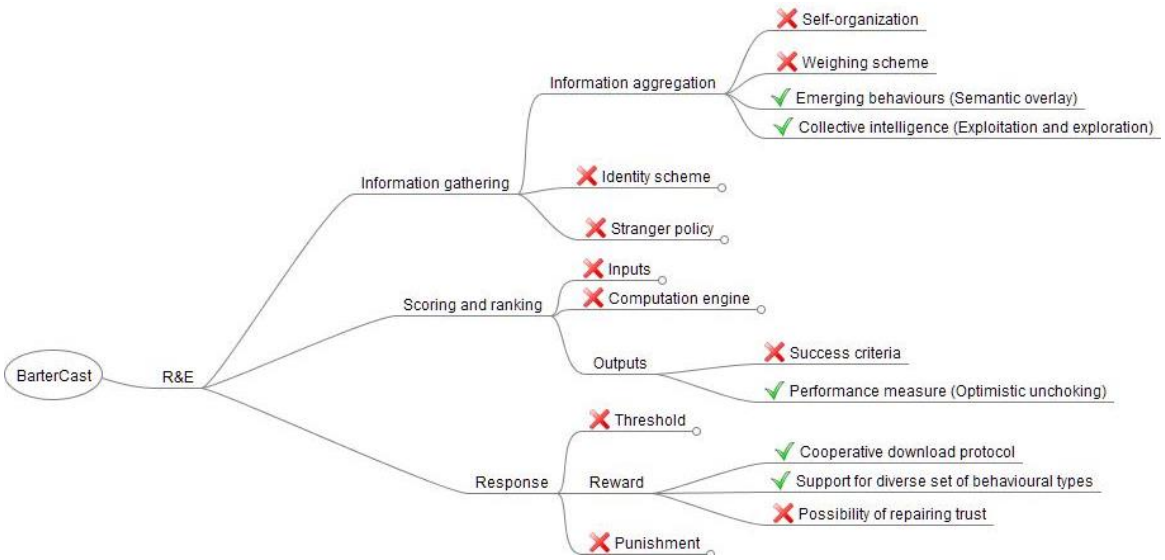


Figure 10: *Evaluation of the interdependencies between resilience and evolutionism and BarterCast's subcomponents.*

Time and System Dynamics (Figure 11). Coping with dynamics in BC is mainly done on a network level, through availability and connectivity checks, by considering the node churn in the network. Also, many time properties that are important for achieving consistency in operation among the components are not taken into account. Although the validity of the reputation information is based on the 10 most recent transactions, this choice is fixed rather than based on an adaptive policy of system or interaction dynamics.

One possibility that may account for the timeliness of reputation information to be included in this RM is to introduce discounting with respect to information recentness and the dynamics of the subsystem it appears in. Therefore, another useful metric for evaluating RM dynamics is the rate of change of parameters like context switching, resource/service utilization, change of behavioural type, reputation fluctuation, etc. These can all help in tuning the frequency with which the actions and design decisions for a given subsystem will be evaluated, changed or taken into consideration in the first place.

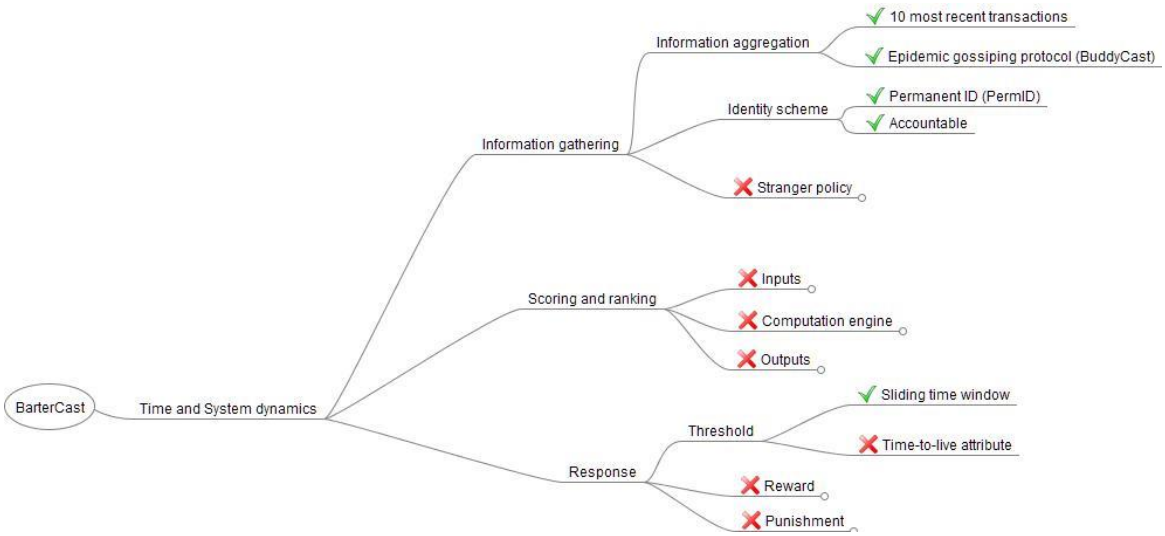


Figure 11: *Evaluation of the interdependencies between time and system dynamics and BarterCast's subcomponents.*

3.4.2 Yahoo! Answers

This section provides a brief overview of the design of Yahoo! Answers' and its mapping onto the framework. For a broader description of its design, see Zaltzman and Le (2007).

Yahoo! Answers (henceforth denoted as Y!A) is a social platform that provides an interface for a large user community to pose or answer questions posed by others (Zaltzman and Le, 2007). Despite its centralized architecture, Y!A is not concerned with providing objective and verifiable information. Rather, its goal is to encourage participation from a variety of contributors, with no commercial purpose for enabling transfers of goods for monetary reward. It is highly hierarchical, comprised of RMs calculating several types of reputation called *karma*. Y!A's design premise is to identify bad actors, give support to honest contributors and remove the load from the customer care team. The latter results from the fact that systems for dealing with abuse that rely on a heavy human-operator intervention are extremely expensive and not necessarily efficient and effective.

Information Gathering. The answers on Y!A are subjective. It is not the goal of the moderation system to distinguish right answers from wrong or place any importance on the objective truth of answers. Rather, it is the distributed effort of the community that determines which responses are ultimately acceptable. All users have two data records that can hold and supply information for reputation calculations: a global Yahoo! user record, which includes fields for items such as registration data and connection information, and a record for Y!A, which stores only application-specific information. Y!A's reputation system considers a few different classifications of users: *authors* create the items (questions and answers) that the community can moderate; *reporters* determine that an item (a question or an answer) breaks the rules and should be removed; *customer care staff* evaluate appeals in cases when the community content moderation process is abused (trusted users who abuse their power or simply make a mistake), but the number of such cases is assumed to be much lower than the total number of abuses. The goal of Y!A's trust system is to reduce the staff's participation in the content moderation process as much as possible.

Scoring and Ranking. User contributions are the objects that users make by either adding or evaluating content. Arriving at a high rate, questions are the starting activity point. Answers arrive much faster than questions and represent the bulk of the reputable entities in the application. All answers are associated with a single question and are displayed in chronological order, oldest first. After making several contributions, the users rate the answers with a simple thumbs-up or thumbs-down vote. The author of the question is also allowed to select the best answer and give it a rating on a five-star scale. If he does not select the best answer in the allotted time, the community vote is used to determine the best answer. Users may also mark a question with a star, indicating that the question is a favourite. The policies for the evaluation process are imposed by two documents: the Yahoo! Terms of Service and the Y!A Community Guidelines, which are the written standards for how questions, answers and reporting should be treated on the platform. The most important properties that the Y!A reputation model includes are the various types of *karma*: reporters of abuse karma, author's karma, answerer karma, content item (question and answer) karma and an inferred karma. The latter type is used to infer reasons for trusting a user or being suspicious about a user from the very beginning, before the user even starts to develop a history with the reputation system. These karmas are used to weigh the decisions made by the according entity, or calculate the level of confidence that a certain entity possesses.

Response. Some community members report abuse for altruistic reasons, out of a desire to keep the community clean. Since downplaying the contributions of such users is

critical, the more public their deeds become, the less likely it is that they will continue acting out of altruism. Therefore, Y!A is not based on the transfer of goods for monetary reward. Other community members have egocentric motivations for reporting abuse. The system appeals to those motivations by giving those users an increasingly greater voice in the community. Finally, an important feature of the system is the well-structured scheme for punishing adversaries. The consequences for malicious behaviour could even reach as far as closing a user's account and banning them from using any type of service in the Yahoo! platform in general.

Stranger Policy. Reports by users with unknown reputation, who had never previously reported an item, are all given equal weight, but significantly lower than reports by users with a positive history. In Y!A, individual unknown reporters have less influence on any content item, but the votes of different individuals can accrue quickly. At the same time, the individuals accrue their own reporting histories, so unknown reporters do not stay unknown for a long time. Thus, the reputation system gives equal weight to reports from bad reporters and unknown reporters.

3.4.2.1 Proof of concept

The way Y!A maps to the new taxonomy framework is presented in Figure 12 to Figure 15. As with BC, useful insights for the improvement of Y!A become visible.

Active Entity Behaviour (Figure 12). Because of its widespread adoption, Y!A attracts trolls and spammers trying to compromise the system by flooding it with harmful and abusive content.

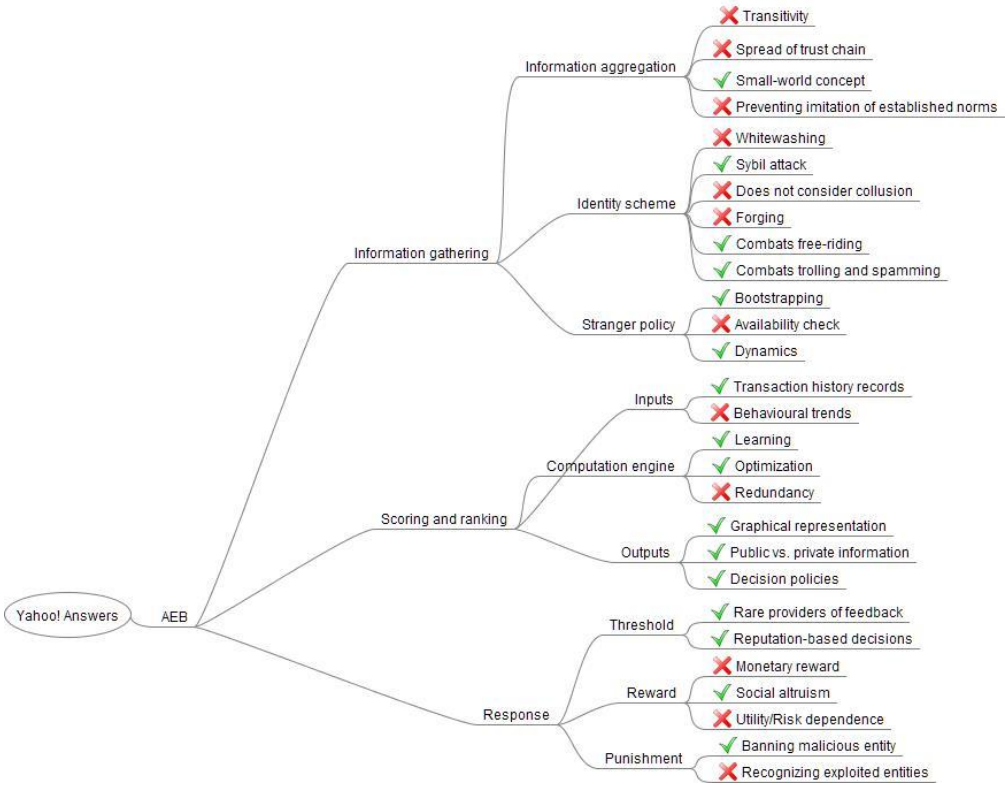


Figure 12: *Evaluation of the interdependencies between active entity behaviour and Yahoo! Answer's subcomponents.*

It is interesting to point out that most of this content was found to originate from a small number of bad actors. According to the *connectivity* property defined earlier that categorized trust systems as densely connected, the overall behaviour and performance of trust systems change when a small number of the entities exhibit change. This implies that

Y!A would have been condemned to failure had it not accounted for the problems caused by these malicious actors. Well-behaving entities would have no incentive to share knowledge of a system that renders their contributions useless in the flood of bad content. Similarly to the "lemon market" (Akerlof, 1970), this inevitably leads to a "sustainable" bad behaviour. One detrimental consequence would be leaving abusive content in high-traffic locations on the site, thereby eroding the perceived value of social interactions over time for the Yahoo! platform in general.



Figure 13: *Evaluation of the interdependencies between context and Yahoo! Answer's subcomponents.*

Context (Figure 13). A significant amount of the information in Y!A is contextualized. Not only do the active system entities have roles (authors, reporters, customer care), but also the passive entities (answers, questions, reports) are categorized according to the topic they cover. The inclusion of an appeals process in the application provides a way to catch false-positive classification by reporters. Nevertheless, by limiting the number of non-reviewed reports to five, Y!A is extremely vulnerable to an automated spam. Furthermore, by adversely colluding to reinforce each other's karma, malicious users can subvert the system by carrying out highly detrimental and low-cost collusion attacks.

One way to partially deal with this issue is to make the number of allowed abuse reports adaptable and dependent on users' (both reporter and content creator) karma. However, this will only prevent colluding behaviour in the context of report abuse, but it might shift the problem towards increased collusion in the context of question/answer karma. For that purpose, it might be worth revisiting the option of marking a certain question as a favourite in the first place (i.e. marking it with a star). Instead, only the number of replies can be taken as a measure of interest in the question and, thus, represent its importance.

Resilience and Evolutionism (Figure 14). One of the design decisions that best complies with the framework principles is the distribution of responsibility (in this case from the customer care team to the users). This implies shifting the intelligence to the connections between the system entities instead of treating it as a component property. Furthermore, Y!A avoids imposing control from outside; instead, the control is mostly distributed across the entities (users), with the necessary intervention from outside (customer care). It is worth noting that customer care agents also have a reputation – determined from the accuracy of the appeal results, although this reputation is not used by this trust system. Nevertheless, the fact that such a score exists for the centralized components of the system works in favour of avoiding trust "institutionalization" in a centralized system.

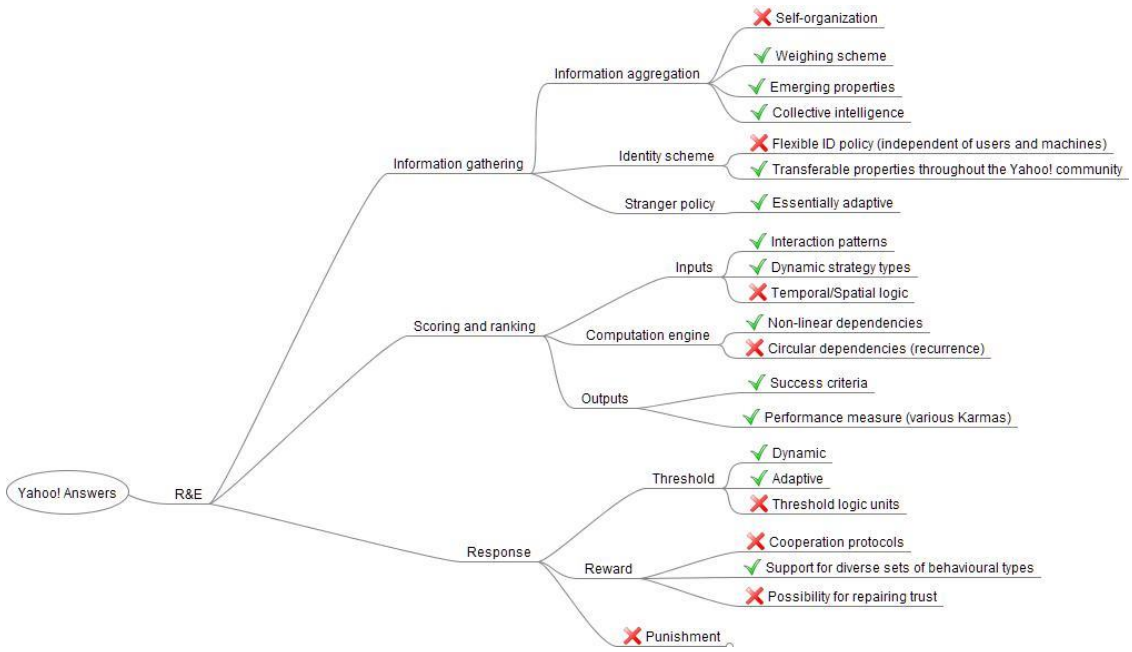


Figure 14: *Evaluation of the interdependencies between resilience and evolutionism and Yahoo! Answer's subcomponents.*

Time and System Dynamics (Figure 15). One of Y!A's design goals is coping with the dynamics, the varying entity behaviour, the ever-increasing amount of content and the uncertain quality of users' contributions. Although Y!A integrates several types of threshold at points requiring decision-making (for example, classifying content as abusive, categorizing new users as suspicious, etc.), these thresholds are predetermined rather than adaptive according to the system dynamics. However, there are time-driven events conditioned by the activity/inactivity of users. For example, if a question's creator does not rate some of the answers as the best ones, the same action will be taken based on the community votes. Unlike BarterCast, Y!A implements weighing policies to give different importance to entities' actions, based on their reputation. Nevertheless, the trust system gives equal weight to reports from bad reporters and unknown reporters. Although such a property is justifiable if a negative karma is to be publicly displayed, this could lead to leveraging certain types of malicious behaviour, such as whitewashing, and will let abusive behaviour flourish unnoticed over a long period of time. Therefore, the RM should (at least) internally use the score from bad reporters to decide about their rights, privileges and actions in general.

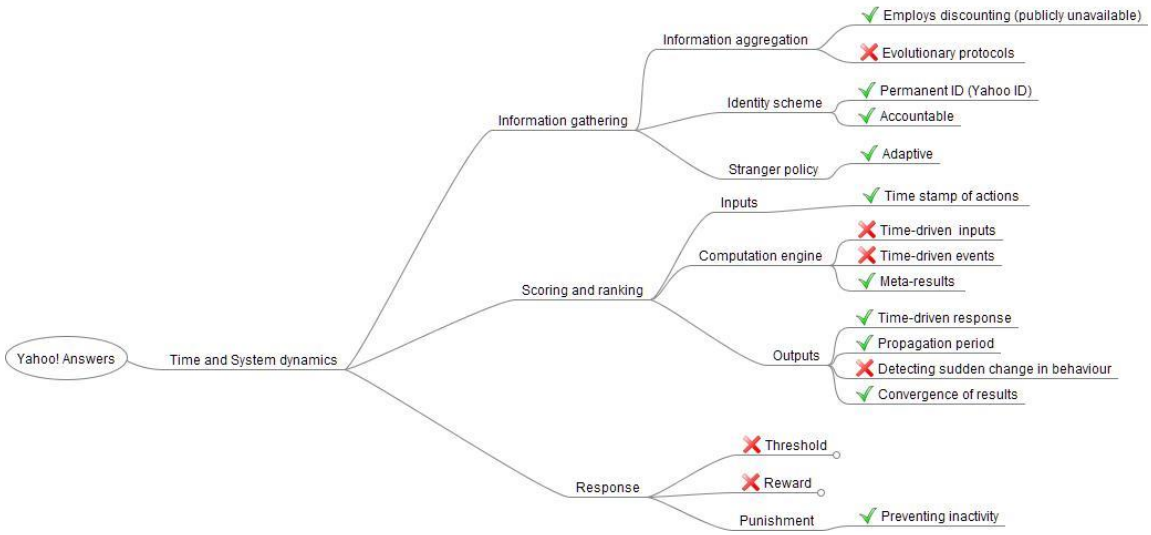


Figure 15: *Evaluation of the interdependencies between time and system dynamics and Yahoo! Answer's subcomponents.*

3.4.3 Summary and Discussion

The power of case study research facilitated the development of a taxonomy framework and the evaluation of the design of two cases: BarterCast and Yahoo! Answers. The analysis of both systems and their mapping onto the taxonomy framework demonstrated the consistency of the systems with GST in a direct and efficient manner. This, on the one hand, enabled the detection of the points for possible improvements, and the ways of incorporating these improvements. For BarterCast, these apply to the weighting scheme, the credibility assessment of the gathered statistics, the need for more costly procedures for resource utilization, setting thresholds at points requiring decision-making, considering procedures for using the system's implicit knowledge, and accounts for the resilience of the system under various types of attacks. The analysis of Yahoo! Answers, on the other hand, revealed properties of a slightly different nature that required attention. These properties include: accounting for the dynamicity of entities' behaviour, weighting the entities' actions by their reputation, accounting for a wider set of possible attacks that can appear in different subsystems and phases of computing the reputation score, then considering automated spam, collusion and whitewashing. On the other hand, the approach enabled the validity of the developed framework to be established as a conceptual model and a means for guiding the design of human-centric trust systems.

One important general observation is that, being a more socially oriented platform, Y!A accounts for many of the emergent systemic properties, unlike BC, which functions on a more technical level. This is expected and perfectly compliant with the organismic and mechanistic properties of trust systems that were introduced earlier in the thesis; the influence of the human factor in the former (to which Y!A belongs) is much more stressed than in the later (to which most of the BC properties can be ascribed). However, it is important to note that each case (not only the two analysed here) is context-specific and should be evaluated regarding its own design premises rather than compared directly to other cases.

Another observation that concerns both of the analysed trust systems is their lack of consideration in the design phase of the possibility of encountering commercial interests that may be contradictory to the design objectives. Trust and reputation systems have recently become the main facilitators of social interactions and collaboration in the business strategies for information and product diffusion. While this is a normal

integrative part of many marketing policies, it often acts as an opposing force to the processes of information gathering and aggregation, as it creates circumstances for many types of bias phenomena to flourish in the system. For instance, the bandwagon effect (or herd behaviour) of a group of people would suit perfectly a company or any entity that is trying to promote an idea or sell a product, but it may create an overly inflated aggregated rating for the product that will only reinforce the bias and create a whole cascade of such behaviour. Complementing this with the well-known Pollyanna effect (i.e. positivity bias) and the people's reluctance to leave negative feedback gives rise to conditions that will ultimately render the trust system useless. Thus, a certain degree of independence and diversity of entities' opinions and actions must be ensured for a healthy-functioning trust system. For that purpose, a greater diversity of policies (data collection and retention, processing of implicit vs. processing of private information, incentives and punishments, etc.) must take place to capture the contextual and emergent nature of the trust phenomena.

Both of the systems pay little attention to the impact of the community structure on the dynamics of the trust and reputation phenomena. The operation of online trust systems depends highly on the collective actions of the individual entities (users, agents, consumers, sellers, distributors, etc.). Trust phenomena, thus, represent the collective result of the entities' motivations and actions, which, although combined by simple behavioural rules, are manifested in a manner that produces surprising patterns of global and group behaviour. Moreover, these aggregate patterns are fed back to affect the individuals' choices. As entities can make trusting decisions based on their acquaintances' advice or their social network's recommendations, this then affects the diffusion of products and ideas, and in turn influences the dominance of an entity (be it a brand or an identity) in the trust system. Also, the predominant entities affect individuals' decisions about which products to purchase or which ideas to adopt. The formation and the effects of these feedback loops are still seen as a curiosity by many of the trust scholars, although they have been the basis of control theory and systems thinking for a long time. This, together with the low account for the human factor render almost obsolete the ongoing debate on the robustness and resilience of trust systems.

All of the previous observations are mainly due to the global trend of addressing technical and social issues in a mechanistic manner – considering the system merely as a sum of its components and neglecting the interactions between those components. On the other hand, there is another, even more pronounced trend of merging together aspects of systems of a different nature into a single operational whole. The purpose is to produce human-centric systems that would integrate the functionality of the emerging technologies with the self-organizing nature of human societies. In its most obvious manifestation, this trend can be viewed in the advent of social networks, electronic services, e-commerce systems and smart cities, all coming under the umbrella of the Internet of Services and Things. Although both of the trust systems analysed in this work follow this latter trend, it is also the case that most of the other trust systems do as well. Therefore, it is crucial not only for revisiting the system design, but also for redesigning the human-centric systems (to which trust systems belong) as *organismic* in nature, unlike their purely technical predecessors. The failure to recognize this prevents any useful debate on the robustness and efficiency of these systems, let alone the predictability of their entities' actions. For this idea to be conceived and nourished adequately, the research on human-centric systems requires a multidisciplinary conglomeration of models and ideas that will both inspire novel directions and provide the empirical evidence of their sustainability.

3.5 Conclusion and Future Work

Acquiring reputation by means of earning trust from the community is primarily a social process. Online environments can largely benefit from the trusting choices that system entities make during their interactions. The possibility of realizing numerous transactions and virtual experiences in a short time requires scalable solutions for establishing trust. In the modern online dynamic environment, having no trust system to adequately capture the system evolution is equal to being well equipped for a world that no longer exists. The framework presented in this chapter is a systemic approach for designing dynamic, densely connected and environment-oriented trust systems. It improves the existing work by defining four design factors and determining their relation to each of the trust system components. *Active entity behaviour*, *context*, *resilience and evolutionism*, and *time and system dynamics* are defined as major influencing factors of trust system design. These insights are incorporated into a multidimensional framework together with the trust subsystems in order to establish binding connections and dependencies between them. Moreover, they place users at the core of the trust systems. The framework consists of a series of taxonomical trees that introduce granular categorization of the trust system properties. It offers a broader representation of the design choices, but moreover it introduces the notion of emergent properties of trust systems and locates the points where such properties may work against the designer's intuition. Finally, by mapping BarterCast and Yahoo! Answers as representative case studies of socially inspired trust systems onto the framework, some weaknesses were revealed and improvements were proposed accordingly. The methodology of employing case study research to evaluate the approach and to enable the choice of use cases in a non-random manner is rooted in the eminent works of scholars in case study research. By building on and evaluating the theory from case studies, guided by the premises of general systems theory, the presented study provides a blend of social and computational principles that respond to the challenges of introducing systems thinking into the research of computational trust.

Incorporating human elements into the technical workings of online trust systems leads to the emergence of complex patterns of group behaviour that are not necessarily a product of the rules that govern individuals' behaviour. The high discrepancy between the designers'/users' expectations and the trust system performance still remains and is resembled by different forms of undesired behaviour manifested in the collected knowledge obtained from the trust systems. It is important to note that such results do not come from erroneous or malicious inputs from individual entities. The whole chain of dependencies in the trust system has to be considered to determine the causal loops that appear between the system inputs and outputs in order to prevent reinforcement of these phenomena. The presented work revealed many additional interdependencies among the social phenomena present in trust systems, the technical components of those systems and the behaviour of the trusting entities.

A major part of future work should be research on the community structure of the trust networks that account for the quality of relationships between the system entities (both positive and negative). These insights need to be employed in future work that is aimed at studying the convergence of these relationship types in various social contexts in order to model the diffusion of trust opinions and beliefs. Such a model would help to evaluate the validity of the social hypothesis and its claimed computational analogues. Furthermore, the systemic principles elaborated throughout this work will contribute to the realization of any further ideas in the debate about the resilience, robustness and adaptive capacity of human-centric trust systems. Finally, the need for a multidisciplinary approach has become essential for limiting or extending the possibilities offered by information and communication science for preserving practicality, but adding social innovation as well.

4 User Bias Identification in Online Trust Systems

4.1 Introduction

The mechanistic view on computational systems is rooted in the engineering contexts in which a system's behaviour could be predicted by knowing the behaviour of the system's components and their operational ranges. Such systems consist of many devices with known physical characteristics and specifications, and with predetermined communication protocols, such as factory plants or wireless sensor networks. This becomes an issue when transfer is attempted to the online computational human-centric systems for reasons explained in the previous chapter.

Computational trust systems are based on the output provided by the collective of many individual entities (consumers, sellers, distributors, ordinary users). Their motivations and actions, although combined by simple behavioural rules, produce surprising patterns of global and group behaviour. These aggregate patterns also come as feedback that can affect individuals' choices. For example, consumers make buying decisions based on their acquaintances' advice or their social network's recommendations. This also affects the diffusion of products and ideas, and in turn influences the dominance of a brand in a market. The predominant brands then also affect the individual's decisions as to which product to purchase or which idea to adopt. It is therefore important for users to understand that their feedback to the online trust systems they are part of also serves their own purpose.

This chapter provides more insights into the workings of online trust systems and the various ways in which they are designed and function. The users are placed at the core of the trust system and the behaviours that can affect this bonding are analysed through the manifestation of user biases defined by prospect theory. The analysis is performed on a real case, i.e. a practically implemented platform; the results refer to concrete issues, and the proposed solutions concern concrete entities and components of the analysed trust system. A framework for user bias detection and identification is then extracted in the form of a choice architecture for online trust systems. Thus, the chapter tackles and responds to the second hypothesis outlined in the problem statement and demonstrates that it is possible to empirically detect the presence and the type of user bias in a trust system.

4.2 Basic Concepts and Related Work

4.2.1 Online Representatives of Trust

Due to the relatively interchangeable use of trust and reputation by the research community, it is essential that the difference between the two is well understood.

Reputation is the general standing of the community regarding an entity's trustworthiness, based on the past behaviour, performance or quality of service of that entity, in a specific context, i.e. a domain of interest. It is considered a computationally viable representation of an entity's "past behaviour/performance".

Definition 9. A system that facilitates the process of calculating and managing

reputation is called a “reputation system”.

Reputation is the amount of context-aware trust, i.e. a quantitative representation of the trust that the society places in an entity in a certain domain of interest. Trust, on the other hand, is a subjective perception of the trustworthiness of a certain entity.

In addition, the rating systems and the recommendation systems are viewed as online representatives of the trust systems. The rating systems manage the evaluation or assessment of an entity, in terms of quality, quantity or a combination of both. Recommender systems are a subclass of information filtering systems that seek to predict the preference that a user would give to an item or a social element they had not yet considered. This is done with a model built from the features of an item (content-based) or the user’s social environment (collaborative filtering) (Ricci, Rokach and Shapira, 2011).

The paradox that arises in trust systems is based on user preferences, which are overly biased and thus affect their performance and reliability. The preferences themselves are largely influenced by the ratings and the results provided by the trust systems. In other words, there are feedback loops that create reinforcement in the users’ decision mechanisms that amplifies the effect of bias. User bias is conceptualized by the following definition, which has its origin in prospect theory:

Definition 10. Cognitive bias describes a replicable pattern in perceptual distortion, inaccurate judgment or illogical interpretation (Ariely, 2008; Kahneman, 2003; Tversky and Kahneman, 1974).

Trust systems rely to a great extent on the inputs from users about their preferences. Bias in these inputs may have a cascading error effect on the performance of the computational algorithms of the trust systems. As a consequence, not only is the accuracy of the results affected by the user bias, but the perceived system reliability is also directly influenced. Although malleable user preferences affect system performance, they are also greatly influenced by the output from the trust systems. Yet biased behaviour, its causes and effects remain relatively unexplored in the context of online trust systems.

4.2.2 Related Work

Neuroscience has revealed that emotions and cognition, present in different areas of the human brain, interfere with each other in decision-making, often resulting in a primacy of emotions over reasoning (MacMullen, 2003). A very similar, although deceptively simple idea stands behind the work in perceptual control theory: human perceptions are the only reality we can know, and the purpose of all human actions is to control the state of this perceived world (Powers, 2005). The psychology of making trust-related choices is directly related to how people think and feel, perceive and decide. The brain has developed complex mechanisms for dealing with threats and risks. Understanding how it works and why it fails is critical for understanding the causal loop between trust-related perceptions and trust-related choices. Biases can thus be noticed in both the people’s perceptions and in their actions. They arise from various processes, such as: information-processing short cuts (heuristics) and mental noise (Kahneman and Tversky, 1972), the mind’s limited information-processing capacity (Simon, 1955), emotional and moral motivations (Pfister and Bohm, 2008), social influence (Wang, Simons and Brédart, 2001), etc.

A study that is directly concerned with finding the bias and “prestige” of nodes in networks based on trust scores was presented in Mishra and Bhattacharya (2011). One assumption made by the authors is that the macro effects of bias and trust are a direct result of the micro behaviour of the individual entities. This assumption is too strong a

generalization and oversimplifies the relation between the users and the trust system. Another drawback of the study is that it considers certainty and confidence to imply trustworthiness and reputation. Furthermore, the authors do not provide a measure for confidence, and the model is vulnerable to sudden change in entities' behaviour. Dynamics is not taken into account and independent opinions are implicitly assumed.

Although only a narrow set of cognitive biases has been tackled by the research community, some studies have been conducted in this regard (Chevalier and Mayzlin, 2006; Resnick and Zeckhauser, 2002). The authors investigated the so-called self-selection bias, where users only rate the items (movies) they like most, causing an extremely high average rating for the preferred items. Trust systems were proven to be inherently biased towards better-than-average scores. According to Kramer (2007), this is due to self-selection. The phrase "selection bias" most often refers to the distortion of a statistical analysis, resulting from the method of collecting samples. Similarly, in trust systems reviewers are drawn disproportionately from the subset of potential consumers that are favourably predisposed toward the resource. The more accurate the ratings, the greater the degree of self-selection, and the faster the ratings diverge towards a biased state. Hu, Zhang and Pavlou (2009) demonstrate the presence of bias in a deployed system and show that the distribution of online product reviews is not normal for almost 100% of the products. While econometric results prove that the online product reviews for the majority of products sold on Amazon do not follow a unimodal (or normal) distribution, Hu et al.'s results suggest that the distribution of online product reviews on Amazon is bimodal, asymmetric and positively skewed (J-shaped distribution), which is a sign of the presence of user bias. As the findings in the study show, the majority of Amazon's product reviews (>60%) are polarized (either 1 star or 5 stars), while less than 40% of the reviews are moderate (2, 3 or 4 stars). In contrast, the majority of the experiment's product reviews (over 90%) are moderate (2 star, 3 star or 4 star), while less than 10% of the experimental reviews are polarized (1-star or 5-star reviews).

The influence of self-selection in online product reviews on users' decisions to purchase products and post reviews was investigate by Yuan, Hong and Pavlou (2012). The role of two self-selection biases was examined: under-reporting and purchasing biases. The results indicated that users exposed to online product reviews with under-reporting and purchasing bias have a significantly lower intention to purchase a product and post a review. The ratings obtained through self-selection are thus representative only of a specific group of users, as they do not truthfully depict the item's general quality. Furthermore, self-selection is not only a transient phenomenon, but also a steady state of the system (Kramer, 2007). Such issues may appear to have a purely economic nature, but they seriously compromise the reliability and performance of the trust systems. The high impact of online reviews on product sales was demonstrated in Chevalier and Mayzlin (2006) by uncovering some of the motives behind companies' efforts to appear competitive on the market. eBay's success is owed to its reputation/feedback system, yet out of the 57% of users who decide to leave feedback, positive feedback is issued by 99% (Resnick and Zeckhauser, 2002). Moreover, a large amount (41%) of users prefer to stay silent instead of leaving negative feedback. A proposal to interpret silence as part of the user feedback was made by Dellarocas and Wood (2008).

A study that shows how an entire belief system shared by a scientific community ultimately evolves from data across all papers within a specific area is presented by Greenberg (2009). This work finds that as much as 95% of all citation paths flow through four review papers by the same research group. This demonstrates the way citations can be used to generate information cascades that result in unfounded authority of claims. The author concludes that network authority emerges through bias and amplification. Although this work presents retrospective analysis of an online system in order to infer

the observed phenomenon, any attempt at prospective modelling is still disregarded as a viable solution, and a more general hypothesis is neither evaluated nor put forward for future reference.

The study of system entities' behaviour is critical in regard to network structure – regardless of whether it is a traditional social system of interconnected entities or a technical large-scale system. In the work of Watts (2002), the author presents a simple model of how global cascades of network patterns can arise in random networks. The main objective of the study was to explore the vulnerability of the interconnected systems resulting from the network of interpersonal relations. The author considers the information that individuals have about the network and explores how it affects the individuals' decisions. Two regimes are identified in which the network is susceptible to very large cascades. The author then concludes: "*Cascades can therefore be regarded as a specific manifestation of the robust yet fragile nature of many complex systems: a system may appear stable for long periods of time and withstand many external shocks (robust), then suddenly and apparently inexplicably exhibit a large cascade (fragile).*" When it comes to online trust systems, it is precisely these phenomena that act as the cause and effect of non-linearity in complex systems, and that are related to the herding and biased behaviour of the system entities.

4.2.3 Prospect Theory

An area with remarkable research about the irrationality, bias and unpredictability of human actions in various circumstances and mindsets is behavioural economy, more specifically the discipline called "prospect theory" (Ariely, 2008; Kahneman, 2003; Kahneman and Tversky, 1972). Prospect theory describes the way people choose between probabilistic alternatives that involve risk, where the probabilities of the outcomes are known. It divides the decision processes into two phases: editing and evaluation. Editing involves the process where users decide which outcomes are almost equivalent, set a reference point and then consider the lesser outcomes as losses and the greater ones as gains. The aim of the editing phase is to alleviate framing and isolation effects stemming from individuals' propensity to isolate consecutive probabilities instead of treating them together. In the subsequent evaluation phase, people behave as if they would compute a value (utility) based on the potential outcomes and their respective probabilities, and then choose the alternative with a higher value (i.e. utility).

The framing effect is a cognitive bias where people react to a particular choice in different ways depending on how it is presented, e.g. as a loss or as a gain. The framing effect has consistently proven to be one of the strongest biases in decision-making. Prospect theory shows that a loss is more significant than the equivalent gain, that a sure gain is favoured over a probabilistic gain, and that a probabilistic loss is preferred to a definite loss. People tend to avoid risk when a positive frame is presented but seek risks when a negative frame is presented. One of the dangers of framing effects is that people are often provided with options within the context of only one of the two frames.

In the context of preferential reasoning, prospect theory shows that users are often unaware of their taste, even for experiences from previously felt outcomes (Ariely et al., 2006). Not only does this reveal that taste is much more subtle than preference, but it shows that preference itself is not a stable property of human reasoning. Experiments on the persistency of user preferences about identical items at different instances of time proved significant fluctuation in repeated preferential choices (Ariely, 2008; Cosley et al., 2003).

Given that the credibility of the online reviews is susceptible to many types of biases, little research has been conducted to investigate whether consumers trust online reviews.

Intuitively, prospect theory seems applicable to the study of bias in trust systems. However, there are several problems with its practical applicability to the case of trust systems: one is the requirement for known probabilities of the decision outcomes and another is the need for people themselves to set the reference point for the losses and gains. This requires a lot of user effort, as it acts as a disincentive to interact with the trust system. Therefore, only the principles for detecting biases and setting an empirical study are borrowed from prospect theory, whereas the framework for designing trust systems with minimum bias is found in another branch of behavioural economy called "choice architecture".

4.3 Choice Architecture

Choice architecture builds on the findings from the field of behavioural economics that has shown that individuals tend to be subject to predictable biases. These biases in turn lead to decision errors. In order to cope with these biases and errors, different ways have been chosen by behavioural scientists to follow the principles of choice architecture. Some of them have focused on defaults, expecting error, understanding mappings (which involves exploring the different ways that information presentation affects option comparisons), giving feedback, structuring complex choices and creating incentives (Savani, 2014). Others have created a typology of choice architecture elements dividing them into two groups: those that structure the set of available choices and those that describe the choices themselves. Examples of tackling a problem by structuring the set of choices include: determining the number of alternatives, integrating decision aids and tools, setting defaults and presenting choices over time. Examples of describing choice options include: partitioning the given options and attributes, and designing appropriate attributes (Johnson et al., 2012).

The main idea of choice architecture is that choice can be influenced in many ways, while still respecting people's freedom of choice. Just like the architect of a building affects the choice of the inhabitants through the placement of doors, hallways, staircases and bathrooms, choice architects can influence choices in many ways: by varying the order of presentation of choice alternatives, the order of attributes and their ease of use, and selecting defaults, to name just a few of the design options available. In other words, whatever way a choice is presented, it will influence the decision of the decision-maker. In that context, everyone, from a parent presenting bedtime options to a child to a government providing pension options to its citizens, influences the choices and is a choice architect.

The way to approach a problem through architecting users' choices is often defined by the types of biases and decision errors made by the users of the system. However, there is a certain line of work to be followed when reasoning in terms of choice architecture. Next, a general framework is presented for defining the choice architecture of a general system.

4.3.1 Reducing Choice Overload

Classical economics and decision theory support the premise that more options improve consumer and user utility (or at least leave it unchanged). The extra time and consideration that each additional choice demands for evaluation in these theories are often left out, outweighing the benefits of having a broader set of options. Behavioural scientists have demonstrated that presenting consumers with many choices often leads to reduced motivation to make a choice and decreased satisfaction with the choices that have been made (Johnson et al., 2012). This is often referred to as "choice overload" (Iyengar

and Lepper, 2000). Choice architecture can be employed to reduce the choice overload by either limiting the alternatives or providing adequate tools to support decision-making. The ideal number of alternatives will depend upon the cognitive effort required to evaluate each option and the heterogeneity of needs and preferences (Johnson et al., 2012).

As consumption decisions increasingly move online, search engines, recommendation, rating and reputation systems are often the tools that consumers rely upon to find, evaluate and choose products and services. These types of search and decision aids reduce the time and effort associated with information search and aggregation. Moreover, they have the power to shape users' decisions depending on what products are presented, the context of the presentation, and the way that they are ranked, ordered, and presented to the users. For example, research on consumer goods such as wine has shown that the expansion of online retailing has made it simpler for consumers to gather information on products and compare alternatives, making them more responsive to price and quality information (Lynch and Ariely, 2000).

4.3.2 Defaults

A **default** is defined as a choice frame in which one selection is preselected so that individuals must take active steps to select another option (Brown and Krishna, 2005). A large body of research has shown that consumers are more likely to choose options that are presented as a default than to switch to another selection (Smith, Goldstein and Johnson, 2013). Several mechanisms have been proposed to explain the influence of defaults. For example, individuals may interpret defaults as system designers' recommendations. Furthermore, cognitive biases related to loss aversion might be present in the system, or users may simply fail to opt out of the default due to associated effort (Smith et al., 2013). It is important to note that these mechanisms are not mutually exclusive and their influence differs across the decision contexts.

A simple type of default, for instance, is when one choice is automatically selected for all users. A forced choice where a product/service is denied until the user makes a proactive selection can also be in place. Choices that are made repeatedly may also be affected by defaults, such as sensory defaults in which the choice is preselected based on previously gathered information about the users (Johnson et al., 2012).

4.3.3 Choice over Time

Choices with outcomes that will be manifested in the future are usually influenced by several biases. For example, individuals tend to be myopic, preferring positive outcomes or fast results in the present at the expense of future outcomes. This may lead to behaviours such as free-riding in the short term at the expense of longer-term resource security. In addition, individual projections about the future also tend to be inaccurate. When the outcome is uncertain the likelihood of desirable outcomes is often overestimated because the future is generally seen optimistically (Shu, 2008).

Some research indicates that there are ways to structure choice architecture in a way that compensates or reduces these biases. For example, improved decision-making was demonstrated by drawing attention to the future outcomes of users' decisions or by emphasizing the second best option (Shu, 2008).

This thesis is not concerned with the implications of choice over time, but their effect on the detection of biases is taken into account in the analysis of the results from the empirical study.

4.3.4 Partitioning Options and Attributes

It has to be mentioned here that the grouping of options and attributes influences the choices that are made. Partitioning of options includes the division of an entity into categories. For example, household expenses can be divided into rent, food, utilities, transportation etc., and products in a store or on a web page can be categorized into clothes, food, shoes, technology, etc. Partitioning of attributes includes the manner in which attributes are grouped together. A label may group several related attributes together (e.g. useful) or list them individually (e.g. short time, little clean-up, low maintenance). The number and type of these categories are important because individuals have a tendency to allocate scarce resources equally across them and are biased by assigning equal probabilities to all events that could occur (Fox and Clemen, 2005). As a result, aggregate consumption can be changed by the number and types of categorizations. For instance, car buyers can be nudged toward more responsible purchases by itemizing practical attributes (fuel mileage, safety, warranty etc.) and aggregating less practical attributes (e.g. speed, radio and design are grouped together as "stylishness").

4.3.5 Avoiding Attribute Overload

Users usually optimally consider all of the entity attributes when deciding between options. However, due to cognitive constraints, they may face similar challenges in weighing many attributes as in cases of evaluating several choices. As a result, the number of attributes can be limited, weighing the cognitive effort required to consider multiple attributes against the value of the improved information requested for decision-making. This presents an additional challenge if the users care about different attributes. Online forms can thus allow users to sort by different attributes and minimize the cognitive effort needed for evaluating many options.

4.3.6 Translating Attributes

Good presentation of information about attributes can also reduce the cognitive effort associated with information processing and reduce errors. This can generally be accomplished by increasing the evaluability and comparability of attributes (Johnson et al., 2012). One example is to convert commonly used metrics into those that users care about. For example, choice architects can influence decisions by adding evaluative labels (e.g. good versus bad or high versus low) to numerical metrics or vice versa, or by changing the scale of a metric (for instance, listing monthly cost versus yearly cost).

Choice architecture is also in accordance with the systemic principles outlined in the taxonomy presented in Chapter 3. As a result, relying on choice architecture leads to the design of a trust system that will obey the principles of general systems theory as well. The important thing to note is that, although improving the design of trust systems can be achieved using choice architecture, the trust systems themselves can be used as part of the choice architecture methodology for improving the general system they are part of.

4.4 Framework for Detecting and Identifying User Bias in Trust Systems

Although there are many biases studied in the research literature, most of them have been analysed and detected in a verbal setting and demonstrated mainly through small-scale experiments for social (and non-virtual) contexts. Very little work has been done to

demonstrate, analyse and tackle the issues related to the cognitive biases in a virtual setting. By employing prospect theory (Kahneman and Tversky, 1979) it is possible to carry out empirical studies aimed at detecting and identifying the biases in an online trust system. Several types of biases are presented and a framework is developed in the form of a choice architecture for online trust systems.

4.4.1 Empirical Study and Findings

In this section, we present an empirical study aimed at revealing the specific types of user bias in trust systems. The empirical study examines different aspects of user bias and seeks the best method for representing user reviews, enabling more accurate and relevant information. The results obtained are used to derive new pointers for identifying and detecting the user bias in trust systems. No relation to the bias assumptions in trust systems design was indicated in the literature prior to this work. This thesis integrates the empirical study in a holistic view that embeds social theories in the design of trust mechanisms in virtual platforms.

The main objective of the empirical study was to investigate the possibility of employing choice architecture for identifying and minimizing the user bias in trust systems. In that regard, studying the consistency of the users' perceptions with the design of the trust systems was necessary. The main questions for which answers were sought during the study were:

- Which descriptive model reflects most adequately the user perception of the numerical rating choice in a trust system?
- How does a difference in the "tone" of the presented choices in the two descriptive models influence the users' decision?
- Which known biases are characteristic of the social contexts that require trust to be represented quantitatively?
- How does the presence/absence of different contextual elements influence user choice, and is this related to the nature (numerical or descriptive) of the attributes?

Design and Methodology: Two methods were used to gather the necessary data for the empirical study:

1. To achieve a good geographical spread of the respondents, better speed of data collection and independence of the participants' opinions, an online survey[4] was the first method chosen.
2. In order to capture some subtleties related to the contextual analysis of decision-making in the trust systems, direct (one-on-one) interviews were also performed. The discrepancy between the preferences of the majority of individuals and the group preference, the difference in the choices made by different users as newly introduced contextual elements, and the difference and interdependence between trustworthiness and acquaintanceship were considered.

The surveys were based on actual (extant) reviews of an HP laptop and were borrowed from the Epinions[5] website.

Both of the surveys asked the potential customers the same question:

"The following texts represent reviews of a product (HP laptop) from Epinions.com. After carefully reading the review, please state how useful the reviews seem to you."

The reviews' contents are given in Appendix 9.1.

4 www.surveymonkey.com
5 www.epinions.com

The surveys differed in the evaluation choices they offered. The results from variations in the offered choices are as follows:

Three groups of responders were asked to rate the reviews for their usefulness. The first group (survey 1) was asked to give a numerical rating on a scale of 1 to 5 (1 = lowest and 5 = highest rating). The second and the third group were offered descriptive evaluation choices. The last two differed by the "tone of voice" used to describe the rating. Among the possible answering choices offered in survey 2, descriptive ratings that contained explicit negative experiences (e.g. Not useful at all) were offered, whereas the rating choices for survey 3 varied only from neutral to positive, without offering a negative choice as a rating possibility. Table 3 summarizes the statistics related to the empirical study. The table provides information about the possible rating options for each of the surveys, and information about the respondents of the surveys. It is important to note here that we do not claim that a numerical feedback of "1" is analogous to the descriptive feedback "Neutral" or "Not useful at all". The numbers 1-5 on the charts given below are only used for shorter and more practical representation.

The demographic of the responders' groups consisted of a population (both female and male) of 86 people, aged 20-50. Its completion required no special technical knowledge, and the interviewed subjects had no difficulty understanding the assignment. The idea of conducting this type of survey was that no one could conclude what the exact purpose of the survey was. It can therefore be claimed that the results of the survey have not been biased towards an expected outcome by the users in any manner.

Table 3: *Statistics about the experimental setting.*

Survey info / Respond. Info.	Type of Survey			
	Survey 1	Survey 2	Survey 3	
	1-5 Num. Feedback	Negative Descriptive	Neutral-Positive Descriptive	Total
Answering choices	1 2 3 4 5	Not useful at all Hardly useful Somewhat useful Quite useful Extremely useful	Neutral Somewhat useful Quite useful Very useful Extremely useful	
Responded	22	22	22	66

The respondents were divided into five groups. Three groups of 22 people were formed for each survey type. Thirty of the 66 survey respondents were also additionally interviewed. The results from these interviews are presented in the final subsection. In addition to the three groups, another 20 users were asked to independently evaluate each review (10 users per review) in order to use their reviews as a baseline for the rest of the respondents' ratings. They expressed their evaluations both numerically, as in survey 1, and descriptively – 10 users as in survey 2 and another 10 as in survey 3.

The main findings from this empirical study are discussed in the next section. Although the experiment was on a relatively small scale, the results are both interesting and insightful, and reveal important points with respect to the given objectives.

4.4.2 Distinction Bias

As a first step in revealing the potential presence of cognitive bias in the experiment, the results from the separately evaluated reviews offline were compared with those obtained from the three surveys, where the two reviews were evaluated together. The goal was to

identify the presence of a *distinction bias*, which is idiosyncratic to contexts where similar attributes are given together for joint evaluation. Distinction bias is manifested as a tendency to view two options as more dissimilar when evaluated simultaneously than when evaluated separately (Hsee, 1998). The presence of distinction bias implies that comparing two options, as was done in the joint evaluation, makes the small differences between the two options appear salient. This bias is often exploited in commerce scenarios when sellers aim to sell a certain product (called *the anchor*) by placing it along with another, *decoy* product that appears to be the worst option when shown together with the anchor (Ariely, 2008).

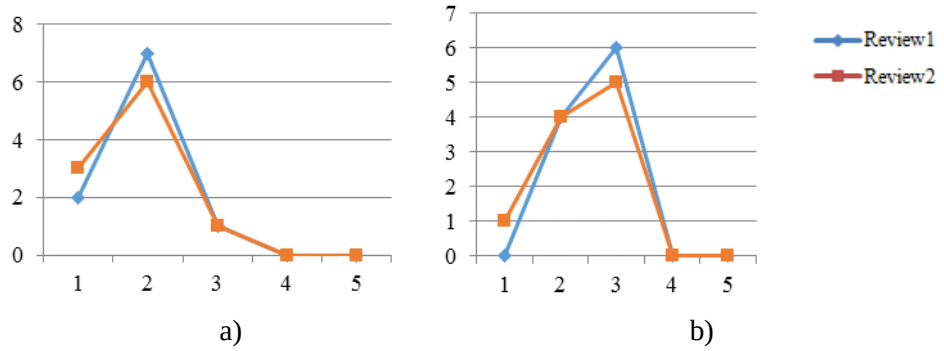


Figure 16: *Rating distribution for each review by offline users a) numerically b) descriptively.*

Figure 16 presents the results from the 20 independent evaluations of the two reviews. The horizontal axis shows the rating values, while the vertical shows the number of users who provided the rating. As shown, both reviews were evaluated as equally useful, 2 on a 1 to 5 scale, and descriptively qualified as “Somewhat useful”.

Figure 17 shows the difference in ratings between the two reviews for each of the surveys. Clearly, they do not match with those shown in Figure 16. In all three cases, review 1 was evaluated as being better than review 2. The mean values for the ratings of review 1 and review 2 for the three surveys are given in Table 4.

Table 4: *Mean value for the ratings of the surveys.*

	1-5 Num. Feedback	Neg. Descriptive	Neutral/Pos. Descriptive
Review 1	3.727	3.545	3.636
Review 2	2.5	3.18	1.95

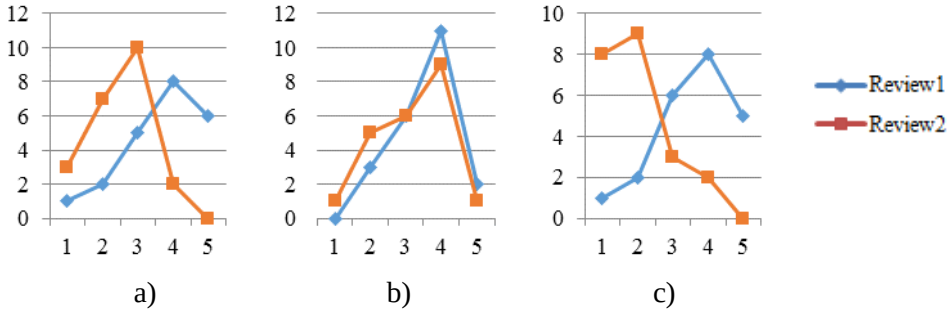


Figure 17: *Rating distribution for the two reviews for a) survey 1, b) survey 2 and c) survey 3.*

Since preferences are often formed through distinction between given options, joint evaluation of recommendations may often result in a choice mismatch. The consequence is that the choice that appeared to be the best option may not provide the best user experience, leading to dissatisfaction.

These results point to the presence of a *distinction bias*, i.e. a tendency to view the two options as more dissimilar when evaluated simultaneously than when evaluated separately. Furthermore, these results demonstrate that there is another form of cognitive bias known as *anchoring*, which describes the human tendency to rely too heavily on a particular piece of information (the "anchor") for which they seek confirmation of their own decisions and use this initial piece of information to make subsequent judgments regardless of the additional information gathered. Thus, there is a significantly increased opportunity to make an error in accurately predicting the utility of the future outcome. The curiosity with this effect lies in the difficulty in avoiding it, even when the decision-makers are informed of the potential for anchoring, and moreover, even if they are offered monetary incentives to avoid it (Simmons, LeBoeuf and Nelson, 2010).

Clearly, it is important to account for the issues of distinction bias and the anchoring effect in the design of trust systems. The question is: what is the cause of those biases, what are their effects, can they be accounted for in practice, and are they just a regular part of any human-centric system? The following sections aim to provide the answers.

4.4.3 Numerical-Descriptive Discrepancy and Positivity Bias

In order to address the effects from the potential presence of distinction bias and the anchoring effect, it is important to realize how they are formed in the particular context of interest. Therefore, it is necessary to understand the differences in the presented evaluation choices and analyse how they affect users' opinions and perception.

Exploration is needed about which of the descriptive surveys is closer to the one with numerical representation of the ratings. The practical implication of such analysis is to examine the extent to which the current trust systems with numerical ratings are calibrated to the users' understanding of those ratings. For this purpose, the rating distributions of the two reviews for survey 1 are compared to those of surveys 2 and 3. The results are shown in Figure 18 and Figure 19 respectively. It can be seen that there is a good match for the distributions of review 1 in both cases. However, compared to the distributions in Figure 16, there is still a great difference between the average rating value for review 1 (1.9) provided by the independent user evaluations, and the average rating value for review 1 in survey 1 (3.727), survey 2 (3.545) and survey 3 (3.636).

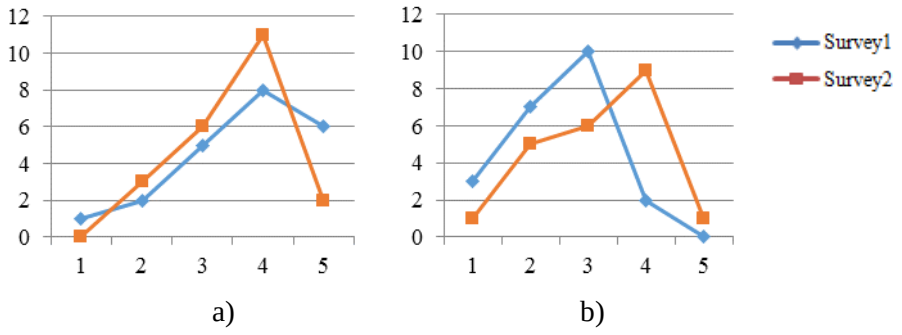
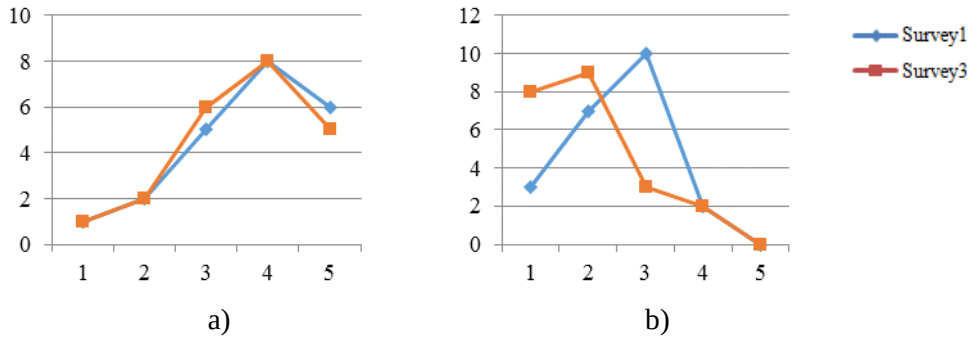


Figure 18: *Difference in rating distribution between surveys 1 and 2 for a) review 1 b) review 2.*

Here, the following causal link is identified: distinction bias is caused when review 1 and review 2 are put together for evaluation, whereby review 1 appears to be the better

option; moreover, review 1 also appears to be “very good” quality, leading to exaggerated positive ratings and, consequently, to positivity bias in the final rating.



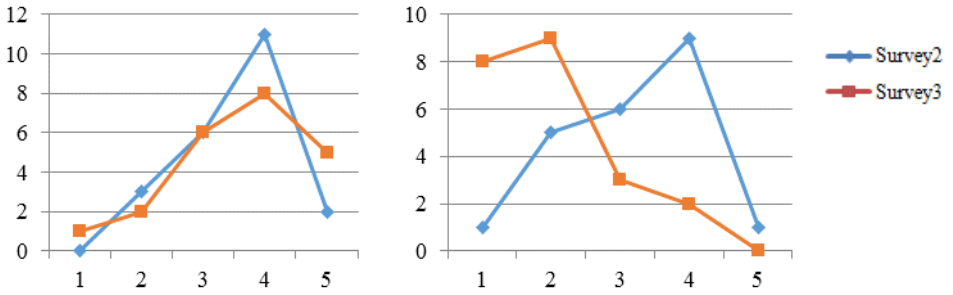
a) b)

Figure 19: *Difference in rating distribution between surveys 1 and 3 for a) review 1 b) review 2.*

The distinction bias in isolation is not the only factor that contributes to such positivity biases. It is human nature to search for or interpret information in a way that confirms one’s beliefs or hypotheses. For example, the *Pollyanna principle* (also known as *Pollyannaism* or *positivity bias*) is the tendency for people to remember pleasant items more accurately than unpleasant ones (Pohl, 2012). Since the user feedback/ratings are given after the user has experienced the item/entity being rated, it is often the case that the given trust score is higher than it should be in order for the users to justify their choice. This is especially the case when the item being rated has already been evaluated positively by most of the previous users. Furthermore, people’s reluctance to leave negative feedback often contributes to overly positive ratings for a certain entity. Such ratings may even seem inflated, although that may not be the case. Thus, it is important to bear in mind that users tend to rate trusting entities relative to the other entities they have perceived in a context similar to the one observed at the time of rating, rather than on a predetermined scale in some mental assessment models.

4.4.4 Positivity Bias and the Framing Effect

In this section, the two surveys with descriptive evaluation choices are analysed. They differ slightly by the tone of positivism, although both offer five choices. Survey 2 contains more explicit negative statements, while the choices in survey 3 vary from neutral to positive. The question to be answered is: *how does a difference in descriptive choices affect users’ evaluation*? To do that, the rating distribution of the reviews for survey 2 and survey 3 are compared. The results are given in Figure 20 a) and b).



a) b)

Figure 20: *Difference in rating distribution between surveys 2 and 3 for a) review 1 b) review 2.*

The results reveal large differences in the evaluations of the reviews between the surveys, although the offered choices differ only slightly. This demonstrates that people tend to draw different conclusions from the same information, depending on how that information is presented, which is known as the *framing effect*. Moreover, it demonstrates that small changes in the input lead to large changes in the output, which is characteristic of non-linear systems (or if we go further, of systems with chaotic behaviour). In Figure 20a), this effect is demonstrated as a slight smoothing of the exaggerated positives. Complementing these results with those shown in Figure 17b), it can be concluded that for survey 2, the ratings for the two reviews also come close to each other. One interesting result is the high mean rating value for review 2 in survey 2. Compared to the results in Figure 16, it can be seen that the presence of a positivity bias is higher when users are invited to choose between negative-positive evaluations, compared to a setting where they are required to choose between neutral-positive ones. Another interesting result is the difference in the number of ratings for the same descriptive value in survey 2 and survey 3. These are shown in Figure 21. The results are shown as survey 2/survey 3. The left chart represents the results for review 1, and the right chart for review 2.

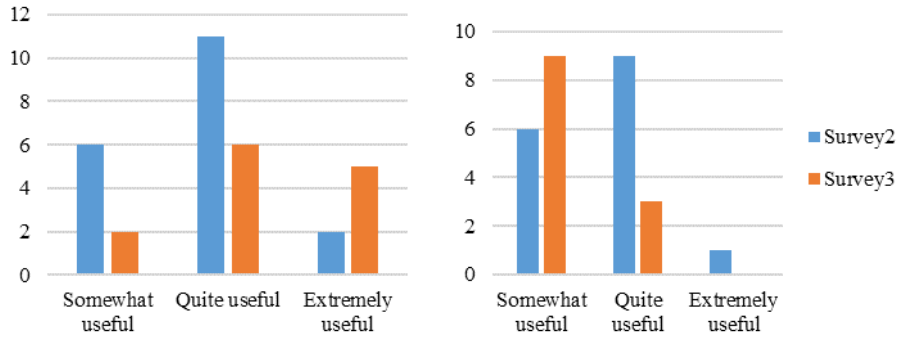


Figure 21: *Difference in ratings between the same options in the two descriptive surveys.*

The representation of the bias is not the only thing to be considered when the user bias is discussed. Biased search and biased interpretation of information are further issues to be taken into account. Users tend to search for information and evidence consistent with their current thoughts and beliefs. Furthermore, two individuals may have the same information, but the way they interpret it can be different, and even differently biased. Finally, there can also be a biased memory, or so-called *selective recall*, when, even in cases when people gather and interpret evidence in a neutral manner, they still remember it selectively to reinforce their expectations.

To reduce or prevent the undesired framing effects, choice should be presented in a variety of frames simultaneously in order for the most accurate and consistent preferences to be determined. If the same problem is presented at the same time, framed in a couple of different ways, it is possible to recognize the impact of framing effects in a more transparent way. Such a strategy of simultaneous presentation might reduce the extent to which a decision-maker is influenced by the framing effects.

These findings enable the issues relevant to biased user behaviour in online trust systems to be accounted for.

4.4.5 Closing the Loop: The market of Lemons on the Market of Opinions?

This section closes the loop of the study on cognitive bias by returning to the point where the discussion started in the context of the online trust system. The investigation of the

impact of cognitive bias required more interactive work with the respondents of the survey. Therefore, 30 people were additionally interviewed face to face. The questions were constructed based on the theoretical considerations presented in the sections above. The answers to the questions were as follows:

Q1. If you know that reviewers will get money (proportionally higher/lower amount, depending on the ratings given) for their reviews, would you give the same grade?

22 answered: *No, one or two grades lower*. **1** was *Not Sure*; **7** answered *Yes*.

Q2*. Imagine that you were a reviewer and that the amount of money you get depends on the amount of money other reviewers (of the same product) get. Would you give the same grade?*

27 answered *Definitely a lower grade*; **3** answered *Still the same grade*.

The three who gave the 2b) answer were additionally asked:

Q3. *If you see that an opinion you consider of low quality is the one that got the highest number of votes, would you reason similarly next time?*

2 answered: *No*; **1** answered *Yes*.

The purpose of these questions was to investigate the reasoning of the respondents as they were required to switch between contexts containing different types of incentives. Although the interview was on a small scale, it provided very interesting information about the ways users make rating decisions. The additional interviews show that the combination of different incentives may lead to different manifestation of the effects from the user biases. Introducing monetary elements must be done with great caution, as it may cause information to be treated as a limited or trading resource in the process of acquiring social capital. This would in turn squeeze the high-quality entities out of the system, leading to what is known as "the market of lemons" on the market of opinions. Therefore, despite the context of use, a framework for designing a trust system with minimum bias must incorporate the knowledge of the entities about the network they are part of, as well as the dependence of the information sources providing the trusting information.

4.5 Discussion

The results obtained from the empirical study that are summarized as exaggerated positive ratings, big differences in the rating behaviour of the users even when a small change in the evaluation options and attributes was introduced, and big impact of the incentives on the rating decisions, imply that the tendency of the users to exhibit cognitive biases is not only the cause of, but also the effect of, the trust system design.

The high number of neutral evaluations demonstrated by the empirical study in Figure 20b), which is even higher than the positive evaluations for review 2 in Figure 20a), confirms the reluctance of users to leave negative feedback. Moreover, it points out the significance of the "Neutral" attribute as a transitional grade between "Negative" and "Positive". Similar to this finding is that the introduction of "Very Useful" as a shade between "Quite Useful" and "Extremely Useful" shifts the exaggerated positive ratings lower on the rating scale. In both of these cases the effects of positivity bias are smoothed. As a consequence, it can be concluded that people's perceptions of the offered evaluation choices are more adequately captured. These findings may also explain why the offline model of social reputation succeeds in detecting untrustworthy individuals without requiring consensus about someone's trustworthiness. The offline world offers numerous opportunities to pick on the hidden clues behind people's intentions. In a similar way, the idea of introducing hidden signals in online trust systems is worth exploring in other contexts as well.

It can be concluded from the above that a methodology for detecting and identifying

biased behaviour must contain the following elements:

- Care must be taken to account for the specific contextual elements: disentangling the collaborative, competitive and monetary elements when deciding on the design of the particular trust system.
- A decision should be made about which of these contextual elements are desirable as public features.
- Adequate representation should be given of the entities' attributes in a way that is consistent with the users' perception of those attributes.
- Proper incentives should be offered for the required user effort.
- An empirical evaluation should be performed to detect the presence of user bias as well as to identify the concrete types of biases in the particular context. This can be done by introducing hidden signals and "shades" of grades about entities' attributes and the trust system options. The aim is to obtain a better and more truthful distribution of the results from the trust system.
- A formal model is required to compute in a technically meaningful way the information gathered by the trust system.

It is also important to note that the proposed methodology does not assume knowledge about the rest of the connections in the network and their quality of interactions. Thus, it offers a possibility of making independent decisions regarding the trustworthiness of the entities of interest.

One question that arises here is: what are the practical implications of these findings for the computational trust models, especially for the computation engines for aggregation of trusting opinions? This is investigated in the next chapter.

4.6 Choice Architecture of Online Trust Systems

Based on the results and the insights obtained from the empirical study, a framework for identifying and coping with user bias in trust systems can now be extracted in the form of choice architecture.

4.6.1 Reducing Choice Overload

The various ways of presenting trust information are called "options" in the context of online trust systems. For instance, a number of stars representing entities' reputation, an icon or smiley face signalling bad/good behaviour, or the contribution/usage resource balance of a community member, etc. Representing trust information in several ways simultaneously can be very useful if the right options are chosen. When faced with a small number of choices, users are pretty good at reviewing each option and then making a choice based on their own preferences. However, the comparison process becomes hard when there are lots of options. Tversky described the process of dealing with a complex set of choices as "elimination by aspects" (Tversky and Kahneman, 1974). This process involves a person deciding on the most important features of the available options and then setting a minimum requirement to make a decision. The purpose is to structure the complex ideas in order to reduce the likelihood of missing out on the correct choice because of the present complexity.

In the context of online trust systems, choice overload happens relatively easily, as trusting information can include several inconsistent representations of the same information. For example, users' trustworthiness can be given as a balance between a user's contribution and usage of resources, in the form of comments and recommendations by other users, or as a reputation score of their general behaviour over time, etc. There is often contradictory information about a user's trustworthiness given by

these representations as each of them tackles a different context. This can be confusing for the decision-maker. As a result, they may opt out of the system, or, as most often happens, they most probably choose what most of the other users have chosen. This leads to herding behaviour, or group bias. Reducing choice overload is thus very important for alleviating such bandwagon effects. Furthermore, accounting for information dependence in case such overload exists is crucial in battling these user biases. Subjective belief calculus has been proven to successfully account for such dependence issues (Cattaneo, 2011).

As Adamic et al. (2011) conjecture, and as the empirical study shows, the influence of hidden signals on users' decisions should not be neglected when reasoning about choice overload. Moreover, it should be considered as a trait of the trust system itself. This implies that it is important to limit the number of similar options given to users. Reducing the choice overload in this manner will also contribute to alleviating the distinction bias in the trust system, while at the same time reducing the system complexity.

4.6.2 Defaults

The simple principle that makes the defaults work is that people take the action that requires the least amount of effort. Good choice architecture work will set the correct default options from the start in order to guide the user in the correct direction with the least amount of action. Thus, proper defaults may also provide an incentive for users to take action, leave feedback and contribute to the trust system.

Some default effects are implied by the situation. In social settings, for example, the normative choice (what others are doing) may be adopted unconsciously as a social default effect (Huh, Vosgerau, and Morewedge, 2014). People are thus more likely to choose what they observe others are choosing, even if they do not believe that the others are more knowledgeable. People are also more likely to treat choices that require less responsibility as defaults.

Once a reputation score becomes part of the users' profiles, the users themselves become identifiable by their reputation. Maintaining a stable identity is important for the reputation value to make sense. On the other hand, being equipped with an online identity as a synonym for one's reputation gives rise to some new dimensions in the context of impression management.

In the empirical study presented, three general contexts in which an online interaction can take place are analysed: one with *pure collaborative* elements, one with *collaborative and competitive* elements, and one with *collaborative, competitive and monetary* elements. It is important to understand which elements need to be adequately combined in order to meet a system design goal.

Once the reputation score is used as a signal for the user behaviour based on which others will make trusting decisions, pure collaborative context in the trust system ceases to exist. The requirement for conducting successful impression management to provide a positive picture for oneself makes the reputation to be recognized part of a user's social capital. This adds a competitive component to the trust system. As demonstrated by the direct interviews that were part of the empirical study, such a context acts as a switch in the users' behaviour. In other words, it triggers actions that would not be carried out otherwise (see Q2 section 4.4.5). Furthermore, the presence of positivity bias implies that inflated reputation values in a system devaluate the reputation itself. Such a situation creates the need for additional incentives that will shift the reference of a good behaviour from competition to collaboration again. What is often done in this regard is to apply monetary incentives to provide the desired behaviour. Mixing pure collaborative context with monetary elements has already been proved to throw a shadow on both the social

intentions and the opportunity for monetary gain (Ariely, 2008). In addition to these more subtle influences, there are more detrimental effects that arise from inadequately combining such context elements. Including monetary elements drastically increases the complexity of the system, and introduces interlocking dependencies between the trust system and the outside world. As a result, it is unreasonable to assume the predictability of the users' and the system's behavior. Failing to recognize this can lead to system degradation, and eventually, system failure.

Examples of unsuccessfully combining collaborative, competitive and monetary elements in trust systems have also been demonstrated in practice. By announcing its Partner Program[6] in May 2007, YouTube explicitly offered to its highest-rated users earning revenue from advertisements placed next to their videos. This instantly triggered a series of events of users blaming each other for using automated programs to inflate their videos' ratings.[7] Similarly, the *New York Times* reported on the case of businesses hiring workers on Mechanical Turk – an Amazon-owned crowdsourcing marketplace – to post fake five-star Yelp reviews on their behalf for as little as 25 cents per review.[8] In 2004, Amazon.ca unintentionally revealed the identities of "anonymous" reviewers, briefly unmasking considerable self-reviewing by book authors.[9] Hence, both theory and practice seem to demonstrate that pure collaborative contexts exist when reputation is used internally in the system (for example, to provide a reference or serve as a regulator for the flow of some system processes). Furthermore, it is worth investigating whether the decision to keep reputation private or expose it as public should be given to the users themselves. Acquiring of reputation may thus not be bound to one's performance, and group biases may be greatly limited in this way.

4.6.3 Choice over Time

A major principle of effective choice architecture is concerned with answering the question "what incentive does the user have to contribute to, in addition to using, the system resources?" Users often search for a pay-off for their actions in the system. Furthermore, when users are presented with a certain choice, they will also need some form of incentive to make the correct/desired choice. A simple and effective incentive in social contexts often comes in the form of making users feel useful. An example of this is showing users that a small change in their behaviour comes at a low cost for them, but has a big positive impact for the community, such as providing feedback in the form of rating and recommendation of products, items or other entities.

The dynamics of trust differs from the dynamics of reputation (Gallos et al., 2012). This discrepancy cannot be captured by any model, as both trust and reputation are intangible matters. Yet, the social models and protocols for detecting malice seem to be successful. In trust systems, one way the dynamics of reputation is embedded in the models is through discounting the relevance of gathered information by some time factor. However, such an approach disregards the importance that some information had in the past in terms of its impact on the outcome. In other words, discounting by recency and frequency is not equal to discounting by impact. Doing the former may be a disincentive for the users who take important actions at a lower rate. Closely related to this issue is the bias of *rosy retrospection*, which is a tendency to rate past events more positively than

6 http://www.youtube.com/creators/

7 http://gigaom.com/video/real-or-robot-the-lisanova-controversy/

8 "A Rave, a Pan, or Just a Fake?" by David Segal, May '11, available at http://www.nytimes.com/2011/05/22/your-money/22haggler.html

9 "Amazon reviewers brought to book" by David Smith, Feb. '04, available at http://www.guardian.co.uk/technology/2004/feb/15/books.booksnews

they were actually rated when the event occurred (Kahneman and Tversky, 1972). This reflects the importance of accounting for the time interval between item consumption and providing feedback about that item.

The experimental study does not tackle any of the concepts related to the dynamics characteristic of trust systems. However, as a disintegrable part of the bigger picture of limiting bias in online trust mechanisms, dynamics must be taken into consideration. Subjective logic accounts for the bias phenomena that arise as a result of the subjective perceptions of the decision-makers by weighing the gathered information according to the time when it was obtained (Jøsang, 2001). Hence, this becomes a requirement for any formal method of processing trusting information aimed at alleviating user bias.

4.6.4 Partitioning Options and Attributes

When leaving feedback on the Internet, raters not only consider the users' own (dis)satisfactions, but they also take into account how a ratee's reputation might be affected by the feedback (Adamic et al., 2011). This fear of bad-mouthing and reciprocation in the context of impression management also translates into a fear of retaliation in e-commerce systems (Resnick and Zeckhauser, 2002). These phenomena explain patterns such as the high reciprocity in public ratings and the absence of negative ratings. Given the reluctance to publicly leave negative feedback, it is reasonable to question the suitability of multi-valued choice in such contexts. Instead, a simple like/dislike option will provide more benefit for all entities that are part of the system. This can help fight Pollyannaism and positivity bias in a similar way to how the shades of grey in the empirical study were shown to shift the rating distribution.

Individual interpretations of trust and friendship vary, and the two concepts are correlated to other characteristics of a social tie and to each other in a non-symmetric way (Adamic et al., 2011). Accounting for the quality of relations in the specific context is thus very important for limiting the user bias in the trust system. The analysis of the quality of relations in a social and computational context has already been elaborated in Chapter 2, where balance theory and subjective logic were introduced. Here, we only pointed out the fact that they hold an important place in the choice architecture for online trust systems.

4.6.5 Avoiding Attribute Overload

While most of the biases mentioned so far were characteristic of an individual, it is important to note that there are also biases coupled with group behaviour. As building reputation is essentially a social process, a system designer should also account for the conditions responsible for the group biases that may arise. Inspecting the properties of each bias independently is certainly important in this regard. However, as was demonstrated by the experimental study, biases are in reality often coupled together, acting as both the cause and the effect of human perceptions and actions. Thus, complementing a few biases together does not mean that their causes and effects will work in an additive fashion.

In the work of Zhang (2011), the anchoring effect that item ratings have on user preferences was studied. Users' inclination toward providing positive feedback were found to be additionally amplified if the users saw the current rating that the item got from the rest of the community. People imitate, or do what others do, especially when having no determined preference. This is often the cause of the so-called group polarization, bandwagon effect or herding behaviour, depending on the field of study (Kahneman, 2003). A study on group polarization on Twitter showed that like-minded

individuals strengthen group identity (Yardi and Boyd, 2010). In other words, when part of group situations, not only are people guided by what the others do, but they also make decisions and form opinions that are more extreme than in individualistic scenarios. For trust systems, this implies that it is not only important to acquire a significant amount of user feedback, but to also investigate whether this amount of user opinions was inferred from a sufficient number of independent sources. A formal apparatus is needed to resolve such issues in a computational setting. An example of that is provided in Chapter 5.

4.6.6 Translating Attributes

Translating attributes is a tool that provides users with the possibility of understanding what they are going to experience when they interact with the trust system, and what the result will be of that interaction. This form of mapping is particularly important for online services where the interactions are usually not face to face.

Prospect theory has demonstrated that people underutilize consensus information, and when given descriptive facts about the quality of a person, they make choices regardless of the statistics offered about that person (Kahneman and Tversky, 1972). Josang and O'Hara (2010) provide a formal proof of this phenomenon known as the *base-rate fallacy*, and give a formal framework based on the subjective logic theory to account for this fallacy in a computational setting. Information offered in trust systems is to some extent a statistic produced by the collective efforts of the community members. This information can be represented in various ways – numerically, descriptively, as a single or multi-valued component, or as a combination of those. Therefore, it is important to explore how users perceive different ways of presenting information, and whether the type of presentation of an entity's trustworthiness collides with the users' perception. The empirical study demonstrated that the trusting decisions (i.e. the ratings given) vary significantly depending on the presentation of information. The best way to represent the entities' attributes depends on the context. However, offering attributes that are compatible with the users' perceptions prevents the rise of user biases. For example, positivity bias and Pollyannaism will be limited if the rating distribution is shifted towards more truthful grades for the entities' qualities. As a result, incentives will be provided for the well-behaving entities to retain their good behaviour without direct punishment for the non-contributing users.

It is important to note that all of the principles of choice architecture are related to each other. For example, proper set of choices will also act as an incentive for users to interact with the system in a meaningful manner. Such actions will lead to truthful feedback results that serve as facilitators of the users' trusting decisions and interactions.

All of the above is further confirmation that trust is a subjective property of the decision-maker and should be treated that way in a computational setting as well. Employing subjective logic comes as the natural reference for the formal apparatus for processing trusting information. With this in mind, a formal method based on subjective logic is presented in the next chapter. This is a confirmation of the second hypothesis outlined in the problem statement of the thesis.

4.7 Concluding Remarks and the Way Forward

Trust is a feeling, a model and a reality, with people at the core. Understanding how they work and how closely they resemble each other is essential for the design practices of online trust systems. The success of trust-based methods depends on the research to identify when and in what manner these methods offer most benefit. This thesis is a contribution to that purpose. The results obtained in the online trust systems will enable

the provision of reliable metrics for building trust and reputation on the Web, as trust awareness is needed on the Internet as it is in the real world.

This chapter has analysed user bias as one of the major causes of malfunctioning in the current trust systems. Several issues were detected and the assumptions were confirmed for the inclusion of bias prevention techniques in trust systems design. The discussion presented shows how these issues affect the work of the trust systems and the users' understanding of them. The relation between trust systems and user behaviour was also explored with respect to the detected issues and concrete types of biases. In an empirical study that included 86 users, it was found that users exhibit distinction bias, positivity bias, anchoring effect and framing effect. This work is a first attempt to offer a holistic interdisciplinary analysis without the mechanistic view typical of engineering approaches. Based on theoretical and statistical analysis, together with the findings of the empirical study, a framework based on choice architecture was proposed for consideration for trust system design. The possibility of employing hidden signals and shades of grey in trust systems for alleviating some of the detected biases was analysed.

The dependence of trust sources and the diversity that can arise from the conflicting opinions of users was discussed as well. Although some current models evaluate trust system performance by comparison with the other existing models in the field, the reference for non-biased results becomes critical when the trade-offs between cost and complexity, usability and robustness are taken into account.

The following chapter concentrates on investigating the application of subjective logic for formal reasoning with human beliefs in trusting interactions with user bias. The main aim is to develop a formal model for limiting or alleviating user bias. The principles of the social and computational theories studied and analysed in the previous chapters are joined to the subjective logic framework and used in formalizing the representation of trusting relationships under biased user behaviour.

5 A Formal Model for Alleviating User Bias in Computational Trust Systems

5.1 Introduction

This thesis as a whole is focused on developing a methodology for designing human-centric online trust systems with limited impact of user bias. In that context, it is important to again stress that the nature of human decision-making in general is unpredictable and highly context-dependent. This implies that complete success in limiting the influence of the human element on computational decision mechanisms cannot be claimed. However, theoretical and statistical analysis provide sufficient evidence for why such an approach would be reasonable and possible.

The previous chapters were centred on defining the design and functioning principles of human-centric trust systems. They also outlined (theoretically and empirically) the importance of the connection between transitivity, subjectivity and bias to the trust system and the results it produces. This chapter includes the application of social and computational theories in a statistical analysis of bias in a trust system. A formal model for trust reasoning based on subjective logic is presented as well. The model inherits the characteristics of subjective logic for the aggregation of subjective human opinions. It is designed to support a trust system that is human-centric, while at the same time preventing or avoiding the present bias.

This chapter addresses several of the hypotheses presented in Chapter 1. The interdependence between bias and information aggregation is presented in a formal model that captures the dependence between bias and trust information. With this, it is possible to account for dependence, transitivity and subjectivity as contextual parameters in the mechanisms for trust aggregation. This model wraps up the main hypothesis and the thesis goals as well, but opens up some novel issues to be addressed in future research. The formal model aims to address all of the three systemic properties of trust systems (dynamicity, environmental orientation and dense connectivity) that were detected and elaborated in the design of the novel taxonomy in Chapter 3.

5.2 Statistical analysis

This section describes the data sets used in the statistical analysis, and provides formal support for the bias reduction method developed later.

5.2.1 Data sets description and statistics

The statistical analysis performed in this thesis relies on the data scraped by Leskovec et al. in the Stanford Network Analysis Project.[10] From there, three large online social networks were taken. They contain labelled links representing negative, positive and neutral sentiments. The data sets come from the Epinions, Slashdot and Wikipedia web portals.

10 All data sets are available at http://snap.stanford.edu

Epinions aims to help people make informed buying decisions. It is a consumer review web platform for product evaluations and recommendations where users are connected into a network of trust and distrust. These are combined with review ratings to determine which reviews are most authoritative. The data set represents a signed network comprised of 119,217 nodes and 841,000 edges, of which 85.0% are positive. A total of 80,668 users received at least one trust or distrust edge, while 49,534 users created at least one and received at least one signed edge. When counting only the triadic closures, 92.9% appear to be positive and 7.1% are negative.

The term "closure" denotes the third (closing) edge from the transitive relation.

Slashdot is a technology-related news website. Summaries of stories and links to news articles are submitted by the site's readers, and each story becomes the topic of a threaded discussion among users. Discussions are moderated by a user-based moderation system, in which users can tag each other as "friends" or "foes". Thus, the network of system entities is signed, and the semantics of a link is similar to that of Epinions: "friend" implies liking another user's comments, and "foe" means disliking another user's comments. The Slashdot data set obtains a network of 79,121 users and 515,581 edges, of which 77.4% are positive. A total of 70,284 users received at least one signed edge, and there are 32,188 users with non-zero in- and out-degree. When analysing only the triadic closures, 87.8% are positive and 12.2% are negative.

Wikipedia is a collectively authored encyclopedia with an active user community. It is based on an openly editable model whose community has developed many policies and guidelines to maintain and improve the site. The data set analysed here corresponds to votes cast by Wikipedia users in elections for promoting individuals to the role of administrator. A signed link indicates a positive or negative vote by one user on the promotion of another (+ for a supporting vote and − for an opposing vote). Using the latest complete dump of Wikipedia page edit history (from January 2008), we extracted all administrator election and vote history data. This gave us 2,794 elections with 103,747 total votes and 7,118 users participating in the elections (either casting a vote or being voted for). Out of this total, 1,235 elections resulted in a successful promotion, while 1,559 elections did not result in the promotion of the candidate. About half of the votes in the data set are by the existing administrators, while the other half comes from ordinary Wikipedia users. The resulting network contains 7,118 nodes and 103,747 edges of which 78.7% are positive. There are 2,794 nodes that received at least one edge and 1,376 users that both received and created signed edges. When analysing only the triadic closures, 84.3% are positive, 11.3% are negative and 4.4% are neutral.

In all networks, the background proportion of positive edges is about the same, with nearly 80% of the edges having a positive sign.

5.2.2 Network Theory Analysis of the Data Sets

The analysis that follows is based on the properties of the transitive triads extracted from the analysed data sets. However, a proof of concept is given as well through statistical analysis. First, network analysis is performed on the three data sets to establish the statistical significance of the triads and study their semantics. A software tool called "Pajek" (De Nooy, Mrvar, and Batagelj, 2012) is used to categorize all the triad types in the data sets. The semantic analysis is done by exploring the network motifs in the data sets using two different tools: FANMOD (Wernicke and Rasche, 2006) and mFinder (Schwöbbermeyer, 2008). With that, the data sets are assigned to known and well-studied families of evolved networks (Milo et al., 2004).

5.2.2.1 Triadic Closure and Transitivity

Triadic census is a way of looking for motifs in a directed graph (De Nooy et al., 2012). In a way, it is a measure of the completeness of the relational triads. An individual's assumption of network closure (i.e. that their friends are also friends) is called transitivity, and it is an outcome of the individual or situational trait of Need for Cognitive Closure (Flynn et al., 2010). Sixteen motifs are commonly defined (Wasserman and Faust, 1994). They are illustrated in Figure 22. These motifs are mainly represented in two ways: by simply numbering the triads from 1 to 16 or by the MAN (mutual, asymmetric, null) model (Holland and Leinhardt, 1974) where: M stands for the number of pairs of vertices connected by bidirected arcs, A is the number of pairs of vertices connected by a single arc and N is the number of unconnected pairs of vertices.

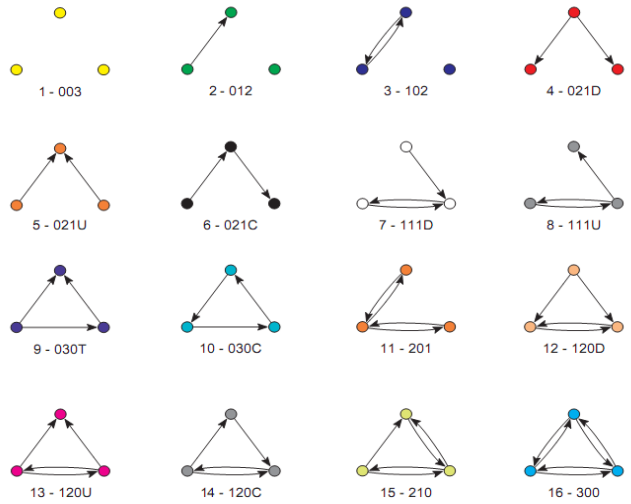


Figure 22: *Representation of all the possible triads that can be formed model.*

When the three numbers are not enough to distinguish among triads, an additional letter is used: D – down, U – up, C – cyclic, T – transitive. For the performed study, only the closed triads in the three data sets were considered. Then, the transitive triads are analysed from the entire set of closed triads.

Table 5: *Closed triads classification.*

Type of triad		
Purely transitive	Partially transitive	Cyclic
(7)9-030T	(11)14-120C	(8)10-030C
(9)12-120D	(12)15-210	
(10)13-120U	(13)16-300	

Table 6: *Number of triads per type.*

	Purely transitive	Partially transitive	Cyclic	Ratio
Epinions	3141472	1763788	61216	2.7
Slashdot	352224	183438	3179	2.9
Wikipedia	566533	35061	6795	14.4

Table 5 shows the division of those types according to the transitivity property of their closure. The term "purely transitive" implies that there is no other type of triadic closure that can be extracted from the triad except a transitive one. "Partially transitive" implies

that are other types of closure that can be extracted from the triad in despite the transitive. The number in the brackets represents the triad type in the triad significance profile of the network motifs (see Figure 23).

Table 6 presents the results from the triadic census analysis obtained from Pajek. It is evident from the results that the ratio of transitive to cyclic triads prevails in all three data sets. This demonstrates the importance of the concept of transitivity of relations in a computational trust system, which was proved long ago by social science in the context of traditional social interactions (Davis, 1967; Holland and Leinhardt, 1974). It also shows the statistical prevalence of the transitive triads over the cyclic ones. Sparse networks are known to contain many null triads (type 003). However, the null triads (which represent isolated entities in a network) are not relevant in the social phenomena rising from the interactions among the community members.

One interesting insight that stands out in the obtained results is the high ratio (14.4) of transitive to cyclic triads in the Wikipedia data set. This is probably because Wikipedia is a relatively closed community, where the requirement for obligatory feedback (vote) by the community contributes to the high interconnectivity of the members. Furthermore, in smaller and more closed communities, there is also a greater social pressure to adhere to the community rules (Schneier, 2012). The network structure of Wikipedia closely resembles the real-world propensity for establishing relationships.

5.2.2.2 Network Motifs

Although the global properties that are exhibited by the networks are important for analysing the overall behaviour of the system, the local structures of various types of networks can demonstrate a wide range of social phenomena that are of equal importance. To investigate the underlying mechanisms of the formation of triads in the computational trust systems, understanding the link between the local (micro) structures of the social interactions between the trust system entities and the global (macro) behaviour of the system is very important. To do that, the context similarity between the data sets and the analysed trust systems must be confirmed. The key method for performing such studies in complex adaptive systems is through finding the triad significance profile (TSP). A TSP is a way to analyse the structure of three-node subgraphs in a directed graph. To calculate the profile of the network, the network is compared to an ensemble of randomized networks with similar degree properties (i.e. a similar ratio of outgoing vs. incoming edges). The comparison to randomized networks compensates for the effects due to the network size. This forms a sound statistical ground for extracting the network motifs.

For each subgraph *i*, the statistical significance is described by the Z score:

$$Z_i = (N_{real_i} - < N_{rand_i} >) / std(N_{rand_i})$$

where N_{real_i} is the number of times the subgraph appears in the network, and $<N_{rand_i}>$ and $std(N_{rand_i})$ are the mean and the standard deviation of its appearances in the randomized network ensemble. For comparison between networks, this vector of values must be normalized to 1. This is important for comparing networks of different sizes, because motifs in large networks tend to display higher *Z* scores than motifs in small networks. Finally, it is important to stress that the TSP of the analysed networks is almost always insensitive to the removal of 30% of the edges or to the addition of 50% new edges at random, demonstrating that it is robust to missing data or random data errors (Milo et al., 2004).

Figure 23 presents the significance profile of the 13 possible directed connected triads in the Slashdot, Wikipedia and Epinions networks of trusting interactions. In Slashdot, the

nodes represent users of the Slashdot services, and the edges represent positive or negative sentiment (friend/foe) directed from one user to another, based on the content that the users had recommended. In Epinions, the nodes and edges are users expressing trust/distrust of one another. For Wikipedia, the nodes are Wikipedia's editors, whereas the edges express promoting/demoting/neutral attitude towards one another. The TSPs of all three data sets are quite similar to each other, suggesting that all these networks may be part of the same superfamily. More interestingly, they resemble something that is between the TSP of the WWW and SOCIAL networks explored by Milo et al. (2004).

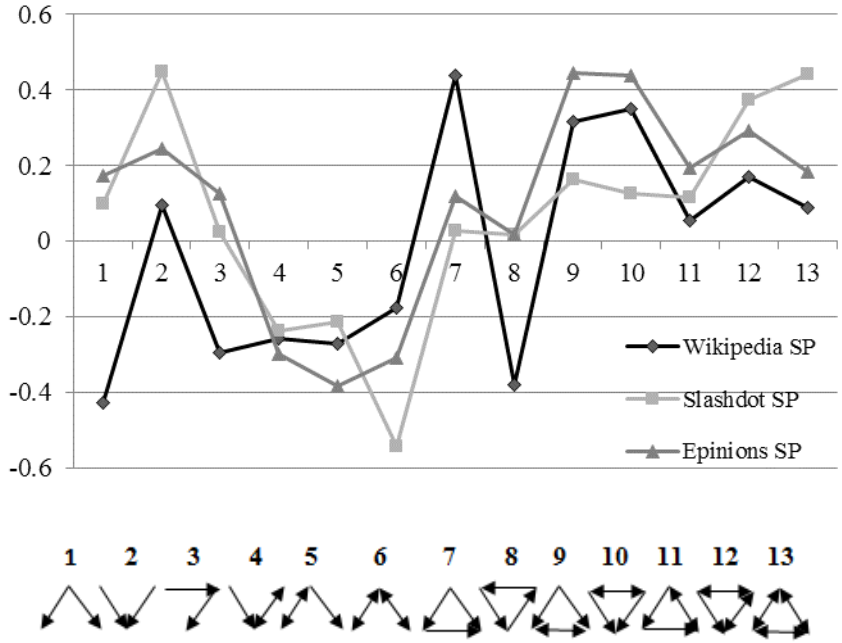


Figure 23: *The data sets triad significance profiles.*

Triads 2, 9, 10, 12 and 13 in Figure 23 have the highest TSP values, while 4, 5 and 6 have the lowest. The types of triads that are present more than the average in the randomized networks are called "over-represented", whereas those that are present less than the average are called "under-represented". The over-represented triads have many transitive triplet interactions (where if $x \rightarrow y$ and $y \rightarrow z$ then $x \rightarrow z$). For example, the over-represented triad 13, termed "clique", has six transitive interactions, the highest transitivity possible in a triad. The less represented triads such as 6, 8 and 11 are highly intransitive.

Wikipedia's triads show a lot of (and much more than Slashdot's) transitivity joined with reciprocity – which was expected, bearing in mind the results from Pajek's triad analysis. It is interesting to note the under-representation of triad 8, which denotes a cycle. However, the two networks also represent different contexts of interactions. Wikipedia is about expertise, while Slashdot is about the ability to recommend good online contents, which is not necessarily a sign of one's expertise in the recommended topic. In addition, Wikipedia's editors were required and obliged to express their opinion (+,- or 0), whereas in Slashdot the feedback depended entirely on the benevolence of users to rate/vote. Finally, the scale of the two networks differs significantly. This is curious, as the Slashdot data set demonstrates neither mere connectedness nor just ordinary social interactions. Also, it is a large-scale data set. The results show that triad 6 is under-represented, which is in line with the expectations as a forbidden triad in social networks (Milo et al., 2004).

Thus, the tendency of social networks to display transitive interactions and transitive

triads is exhibited by the three data sets that are subject of the analysis of this chapter. Moreover, the similarity of the WWW data set in Milo et al. (2004), the traditional social networks and the online trust networks suggests that the classical models of social structural organization may also be used to understand the structure of online social interactions. These findings confirm that understanding triad transitivity is important for studying the link between the micro behaviour of the system's local structures and the system's macro behaviour. Finally, it is important to note that the analyses performed in this thesis are on a larger scale than the social networks originally analysed by Milo et al. (2004).

The analysis only of the transitivity property and the behaviour of the connected triads in the three data sets proceeds with the employment of balance theory, as it enables the implications of user bias in the trust system in general to be extrapolated.

5.2.3 Balance Theory

In this section, the performance of balance theory (BT) as a predictive model of the formation of transitive triads is investigated in relation to the three data sets: Epinions, Wikipedia and Slashdot. The reason why balance theory is the model of choice is that it integrates all of the information properties that are of interest in the thesis: subjectivity, relation sentiments, transitivity. Furthermore, it is often the model of reference in the research literature for the relations formation in the social networks and has rarely been evaluated as a suitable for the given context. With the analysis performed here, its suitability as a predictive tool for the rise of user bias in the trust system is also studied in greater depth. User bias appears to be a great influencer on the predictive power of this social model. Additional issues are also discovered and the design assumptions are generalized to a wider set of social-interaction models. Despite the predictive ability of BT in the context of virtual social networks, its descriptive nature is also discussed.

The presence of user bias in the data sets is statistically identified through the results obtained. Two metrics are discussed that help in the analysis: the accuracy and the Matthews correlation coefficient (MCC). Accuracy is simply the percentage of true classifications out of the total number of classifications. The MCC is a correlation coefficient between the observed and the predicted binary classifications; it returns a value of between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 denotes "no better than random" prediction and −1 indicates total disagreement between prediction and observation. It is defined as:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

TP stands for true positives, which is the number of classifications that were correctly classified as true. TN stands for true negatives, which is the number of classifications that were incorrectly classified as true. FP stands for false positives, which is the number of classifications that were correctly classified as false. FN stands for false negatives, which is the number of classifications that were incorrectly classified as false. The aim of computing the MCC is to establish whether balance theory has the power to be a predictive model of the formation of the transitive triads in an online trust system.

In order to perform statistical analysis of the triads in the data sets, an algorithm was written that extracts only the transitive triads from the data sets. It is written in C++ and has an execution time of less than 30 sec. The pseudo-code and the description of the algorithm is given in Appendix 9.2.1. The results from the computations for each data set are presented and discussed below.

Table 7: *Performance of balance theory as a predictive model*

	True Positive	False Positive	True Negative	False Negative	Accuracy	MCC
Epinions	9646618	211946	567612	535817	**93.18%**	**0.577**
Slashdot	1009165	39169	80901	64097	**91.35%**	**0.565**
Wikipedia	43235	5519	2856	6895	**73%**	**0.191**

Table 8: *Separate sign closure for each triad type in each of the data sets*. The bolded red numbers denote BT's sign prediction for that triadic closure.

	Epinions		Slashdot		Wikipedia	
	% of '+'	% of '-'	% of '+'	% of '-'	% of '+'	% of '-'
(+,+, ?)	**98.7**	1.3	**97.6**	2.40	**92.00**	5.3
(+, -, ?)	54.8	**45.2**	62.2	**37.80**	74.50	**21.3**
(-, +, ?)	41.2	**58.8**	32.7	**67.30**	53.00	**35.7**
(-, -, ?)	**39.0**	61	**53.6**	46.40	**44.90**	46.9

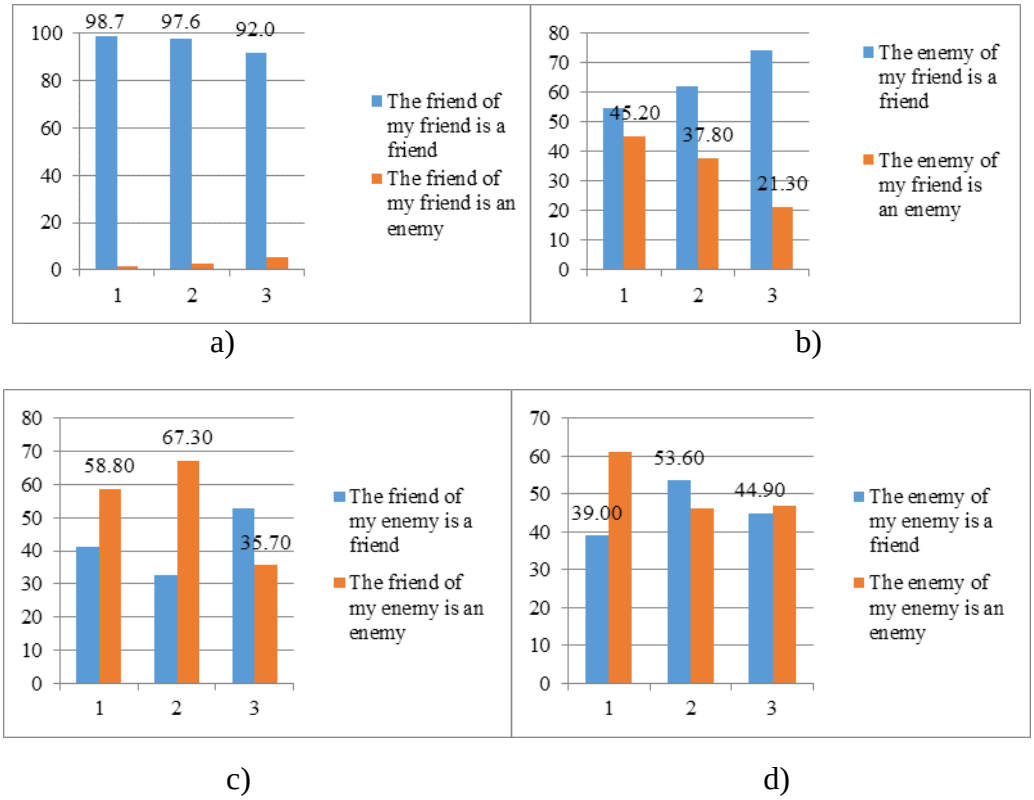


Figure 24: *Balance theory predictive performance on each of the data sets:* 1 – Epinions, 2 – Slashdot, 3 – Wikipedia. According to BT's predictions, the colour that should be dominant on the graphs is: a) blue; b) red; c) red; d) blue.

In Table 7, the accuracy and the MCC of balance theory as a predictive model are shown. It can be seen that BT's accuracy is relatively high for each of the data sets. The MCC, which is a more suitable metric for predictive model fitness, does not show the same (satisfying) performance. These results show that there is a high percentage of true positives. However, this is mainly due to the high initial amount of positive closures in the network, i.e. the (+,+,+) triad, which emerges from the presence of the positivity bias

in the users' ratings and friending affinity. The Matthews correlation coefficient shows that BT's classification performance is at best moderate, regardless of the high accuracy. Performing analysis with balance theory as a social interaction model enables us to realize that the user bias in virtual social networks may not be due to irrational decision-making. Rather, it can come as a reasonable manifestation of users' behaviour in the current design of online trust systems.

5.2.3.1 Implications of Neutrality

Most of the research on virtual social interactions concentrates on applying and testing social theories on unsigned relationships. There is very little research on how neutral (i.e. ambivalent) opinions and ratings affect the performance of the existing trust models. In the empirical study performed in Chapter 4 the presence of particular types of biases was detected in a trust system and the effects of the neutral opinion on the users' decisions was analysed. The obtained results demonstrated that increasing the granularity of the evaluation choices in this way affects significantly the score distribution. More precisely, it contributes to more equally distributed trust scores.

Below is a statistical analysis of how neutral opinion affects the performance of the social and computational models employed in this thesis. The analyses are performed on the Wikipedia data set, which provides sufficient and suitable data for that. Analyses of the separate types of triads are also performed. The results are summarized in Table 9.

Table 9: *Separate triad-type analyses of the triadic closures in Wikipedia's data set* (included are also the ones that contain neutral opinion). Bolded are the results where BT is expected to hold.

		+	-	0	% of '+'	% of '-'	% of '0's
0	(+, +, ?)	39361	2246	1163	**92**	5.3	2.7
1	(+, -, ?)	3251	930	185	74.5	**21.3**	4.2
2	(-, +, ?)	2846	1926	613	53	**35.7**	11.3
3	(-, -, ?)	493	515	91	**44.9**	46.9	8.2
4	(+, 0, ?)	1595	137	146	*84.9*	7.3	7.8
5	(0, +, ?)	1409	541	273	*63.4*	24.3	12.3
6	(0, 0, ?)	69	36	35	49.3	25.7	25
7	(-, 0, ?)	132	85	52	49.1	31.6	19.3
8	(0, -, ?)	176	170	29	46.9	45.3	7.7

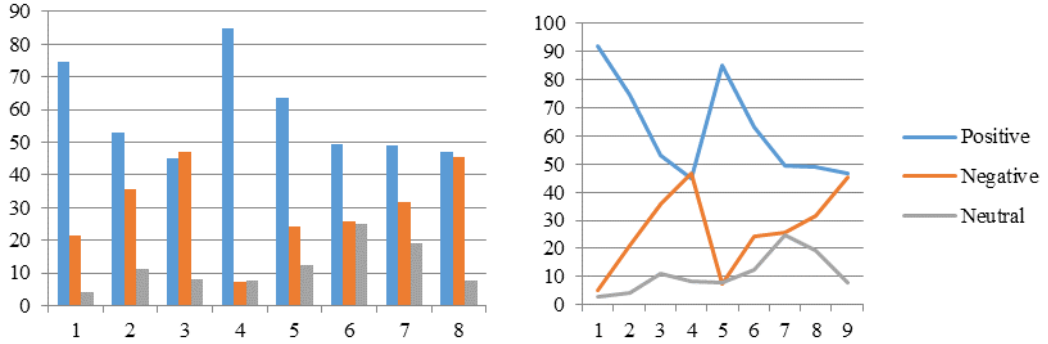


Figure 25: *Separate triad-type analysis of the closures in the Wikipedia data set.* Included in the analysis are also the neutral opinions, and excluded is the (+,+,+) triad for statistical reasons.

Table 9 shows that the only support for BT as a prediction model exists for the triad type (+,+,+), while none of the other triads obeys the BT predictions. It is also evident

that the "neutral" option affects the trust score distribution. This implies that adding a single option for user rating/scoring can have a significant effect on the models for user bias mitigation. Thus, the users' intentions can be more suitably captured by their feedback.

On the bar chart in Figure 25, the x-axis represents the type of triad in Table 9, and the y-axis shows the relative[11] percentage of positive, negative and neutral closures. The (+,+,?) triad is excluded from the bar chart, as it would disproportionally shift the distribution, making the rest of the triads appear statistically insignificant. The reason for this is the high percentage of such triads and the presence of positivity bias.

An important result from these analyses is that there is a notable difference in the tendency for triadic closure between the connections with the semantics "The friend of my enemy is a friend/enemy" and those with the semantics "The enemy of my friend is a friend/enemy". Moreover, this difference is not only more pronounced with the presence of neutral opinions, but balance theory predictions do not hold at all. This is why broader analyses of this issue were performed. Special attention was devoted to investigating this difference and the results of the analysis are presented in Appendix 9.2.2.

Another interesting result is that the greatest shift towards positivity bias happens towards friends' connections. In other words, people tend to show much greater friending affinity towards friends of their friends than towards strangers, even towards strangers about whom the community has a strong positive opinion. Such tendencies are important to consider when designing trust systems, particularly with the choice of a computational engine. These are the points at which a fertile soil for biases is created, as they represent a great source of opinion dependence.

Finally, Figure 25 shows that there is only one type of triad (triad 3) where the percentage of negative closures is slightly higher than the percentage of positive closures, whereas the rest of the triads are inclined towards positive closure. This is to some extent expected, as the overall number of positive ratings indicates that there is a positivity bias in the system.

From the analysis performed, some of the potential problems that may cause balance theory to perform badly as a predictive model can be pointed out. One problem is that balance theory considers the triads (+, -, -) and (-, +, -) as being equally likely to appear. Moreover, it is not concerned with the quality and the probability of their occurrence, or the occurrence of the last of the edges formed in the transitive triad. Thus, not only does balance theory consider "the enemy of my friend is my enemy" as equally likely to appear as "the friend of my enemy is my enemy", it also treats the two as having similar semantics. Clearly, this is not a reasonable assumption.

The effect of neutral opinion (i.e. ambivalence) on balanced relationships is not addressed either in the balance theory model. Furthermore, balance theory conceptualizes the cognitive consistency motive[12] as a drive towards psychological balance, which may not even be possible in a virtual platform with numerous strangers and ad hoc encounters. Finally, the problem with the knowledge about the network that was identified earlier refers to the impossibility of a balanced personal community within the vast and sparse "virtual universe".

In general, the statistical analyses demonstrate that there are important issues that need to be considered in the current theories of social interactions. It is even more important to consider these issues in the current computational theories that rely on balance theory as a predictive model.

---

11 Relative to the highest percentage of the triads shown on the bar chart (1–8), not to the absolute percentage of closures – as we excluded the 0^{th} triad.

12 The consistency motive is the urge to maintain one's values and beliefs over time.

The following section presents theoretical analysis of some of the most important issues that are often disregarded by the research community as a result of the assumptions made about the social interaction models.

5.2.3.2 Knowledge about the Network and the Observer's Problem

Although it is clear that the real and the virtual world are in essence incomparable to one another, it should be taken into account that the virtual world embraces many additional contexts centred on the concept of friendship, similarly to those in the real world. The human-centricity of some of the virtual platforms causes many researchers to take a reductionist approach by directly applying social theories to the computational platforms. Although these approaches are often justified by their practicality, many of them are shadowed by the non-realism in the design assumptions.

This work is also a call for the research community that tackles human-centric and organismic problems in computational systems to be aware of the extent to which the mechanistic view of technical design is appropriate and realistic.

Here, the realism in those assumptions and the implications of their (in)applicability are investigated. This is important to do, as there is a direct connection between user bias and the community structure. As illustrated by Josang (2006), the formation of information cascades and dependent opinions leads to mass hysteria and evokes herding behavior in many contexts. Trust systems being exceptionally fragile to this phenomenon.

Let us analyse one isolated triad from the point of view of each of the entities comprising the triad. Each of the points (A, B, C) in the triad is a different user/node in the network with their own knowledge about the network.

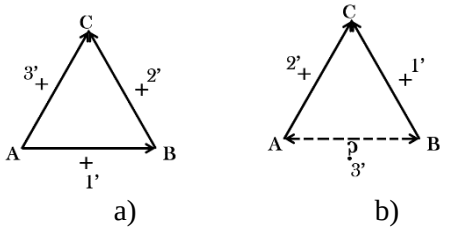


Figure 26: *The (dynamics of the) process of a triad formation.*

This also means that, if we try to look at the triad from A's point of view, we directly assume that A has knowledge about B's intentions towards C, in order to decide about its own friending intentions towards C. Moreover, we assume that there is also knowledge about the existence of that connection in the first place. In other words, there are three implicit assumptions that are often made in the analysis of transitive triads in the literature: first, it is assumed that we are aware of the sequence in which the connections are formed – denoted by numbers as shown in Figure 26 a). Second, it is assumed that we know the friending affinity of each of the entities in the triads towards each other (denoted by the sign of the connections in the same figure). Finally, and most importantly, it is taken for granted that there is some kind of knowledge (or the ability to extract/infer it) about the existence of a relation between any two of the triad entities. These are probably not the only assumptions made in the application of balance theory. We consider them crucial, as without being satisfied, they trigger assumptions that can be rendered useless, undermining the validity of balance theory analysis itself.

But what happens if we disregard these assumptions and concentrate on the realistic side of the practice and philosophy behind the formation of triad connections? Let us start with the assumption that we know the sequence of only two connections, shown in Figure 26 b). The dashed line is the connection whose friending affinity we are trying to decide. In other words, we assume that B→C is the first and A→C is the second connection

established (although it makes no difference). Regardless of the friending affinity (the sign) of B towards C or of A towards C, it is impossible to make an informed decision about the friending affinity of A towards B or vice versa without knowing at least one additional relation (C→B or C→A).

Now, let us disregard for a moment the invalidity of the assumption about a known sequence of relations, and concentrate on the knowledge about the existence of a connection and their nature (sign). The only case for which we would be able to claim some applicability of BT as a prediction model is if we assume that we have the A→B B→C dyad given, and we are trying to decide about the sign of the closing (A→C) edge. However, this is highly unrealistic.

In Lou, Tang, Hopcroft, Fang, and Ding (2013), the authors present a model for triad closure prediction, defining the problem of prediction as follows: given three nodes A, B and C at time t, the model should predict whether A will follow C at time t+1 if it follows B at time t. The context is a Twitter network. In other words, the sequence of relations is B→C, A→B, A→C, which comes down to the previous problem of inferring A→C, knowing the dyad A→B, B→C.

Finally, if we assume that such knowledge exists, it is straightforward to prove that balance theory and the reciprocity in relationships go hand in hand, if BT holds. Taking into consideration the existence of positivity bias in real-world networks and the over-representation of the (+,+,+) triad, it becomes clear why reciprocity and balance theory are often assumed to hold. The problem, however, is multifaceted. First, we assumed that each of the nodes has knowledge about the existence of a relationship and the friending affinity of the connections in the network, which is highly unrealistic. Second, even more unrealistic is the implicit assumption that the users/nodes are aware of the sequence of established connections. Balance theory was built on the assumption of mutuality in the relationships between two individuals, i.e. it did not account for non-reciprocal connections. Therefore, the assumption about knowing the sequence of establishing connection is largely disregarded by most of the works that employ balance theory.

From these analyses, it follows that the only qualification we could give BT is as a descriptive model, which is helpful only for retrospective analysis of existing networks, and not for prospective modelling. The low amount of reciprocity in the social networks analysed in this work, the high amount of positivity bias, and the lack of realism in the initial assumptions contribute to the results obtained being regarded as misleading for the value of balance theory in predicting the social interactions in a network.

5.2.3.3 Opinion Dependence and Conflict

One of the most common assumptions for facilitating the design and analysis of trust systems is the independence of the information sources during the process of information gathering and aggregation. It is very important that each entity in the system independently decides about the interactions and the choices they make in the system based on the information that the system puts at their disposal. However, it is unreal to state source independence on the one hand, and it is impossible to provide it on the other hand, without making privacy violations and increasing significantly the complexity in the system. Being unaware of the implications of opinion dependence can be detrimental to the functioning of the trust system. A situation of opinion dependence is shown in Figure 27 a). Regardless of the signs of the links, nodes A, B and C derive their opinion about X from the same source (D). Moreover, E, who is connected to all three of them, derives its opinion about X from already derived opinions from one source. This can easily be used for inflation or deflation of the general opinion about X when the links are signed relationships, or positive and negative reinforcement in opinions, when the links represent the intensity (numerically) of a relationship.

In addition to opinion dependence, there is another scenario (Figure 27 b)) that is regularly met in the social interactions: this is the case when a user (represented by node A) can get conflicting opinions from their friends about their friending affinity towards another user (in this case node C) for which A requires information. Balance theory or any other social theory does not provide any means for conflict resolution in such cases.

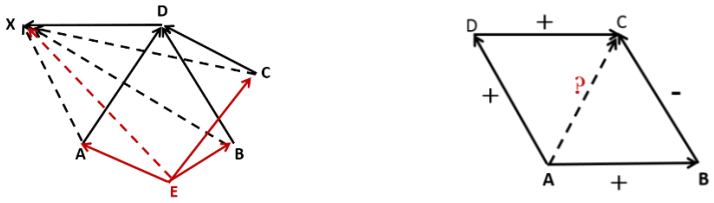


Figure 27: *a) Opinion dependence* *b) opinion conflict.*

Opinion conflict and opinion dependence are closely related to the previous problem of knowledge of the network, to the extent that the two are even contradictory. Whereas the former implies proneness towards considering friends' relations when establishing one's own relations, the latter assumes independence from decisions of this nab Balance theory implies that making friending decisions is based on previous knowledge extracted from other entities.

The main goal of this research was to inspect more closely the relation between user bias and users' perception of trust systems, in order to be able to understand how it affects the performance of the theories of social interactions in the virtual world. Therefore, the following section analyses the possibility of expressing intensity of sentiments and beliefs rather than only a polarized attitude (affinity), and explores the applicability of subjective logic for that purpose. In order to do that, a method is first developed for translating binary ratings to subjective logic opinions. It is then shown that balance theory is only a special case of subjective logic under certain conditions. After establishing the method and verifying its statistical validity, the contributions of the thesis are finalized with the introduction of a formal method for bias reduction that is based on subjective logic. The method is justified theoretically and statistically with the performed studies and analysis.

5.2.4 Subjective Logic

In applying the apparatus of subjective logic, the operators are implemented in real-world data sets and their performance is statistically evaluated. The operators that are employed and evaluated are the transitive trust operators introduced in Chapter 2: the uncertainty favouring operator, the opposite belief favouring operator and the base-rate sensitivity operator.

5.2.4.1 Opinion and Sentiment Representation

In order to define the values of the subjective opinion in the opposite belief favouring operator, balance theory is taken as the baseline. Prior to that, sensitivity analyses of the operators are performed to determine how the values of the resulting score change with small changes in the representing opinion.

All the methods for computing the values of the belief (*b* and *d*), the uncertainty (*u*), and the base rate (*a*), and the other necessary operators were programmed in the MATLAB programming environment. The same data sets from real-world applications were used for the statistical analysis of the subjective logic and its comparability with balance theory: Wikipedia (representing votes for status recognition: 1, 0, -1; context: collaborative writing of encyclopedia articles), Slashdot (friend/foe relations: 1, -1;

context: collaborative sharing of news articles), Epinions (trust relations: 1, -1); context: e-commerce web of trust. It is convenient that the data sets differ in the context, the scale and the value representation, allowing for broader, more diverse and more interesting comparative analysis.

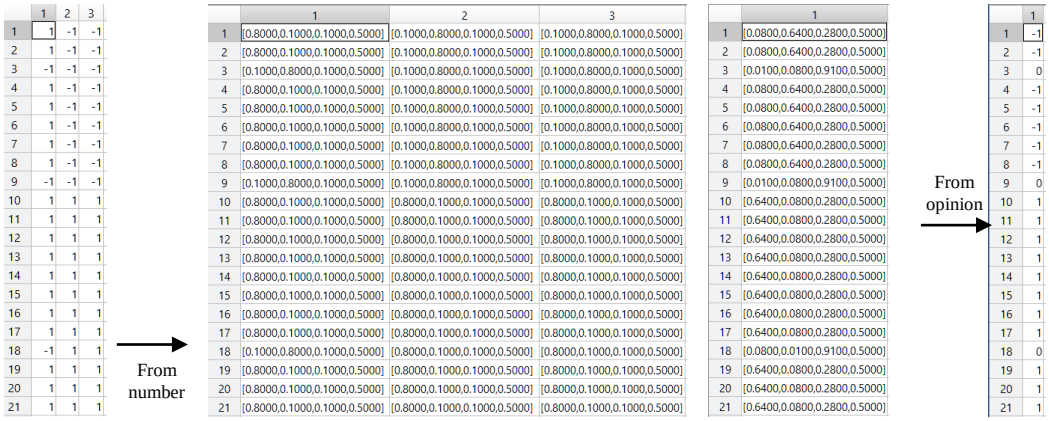


Figure 28: a) *Subjectivization of 1/0/-1 sentiments to opinions; b) desubjectivization of opinions to 1/0/-1 sentiments.*

A method was required for transition from one type of input value (binary, numeric, descriptive, etc.) to a subjective logic representation, and vice versa. An analogy is used from fuzzy logic, where the values are “fuzzified” from and “defuzzified” to “crisp”. The procedure in subjective logic is translated into “subjectivization” (going from “normal” values, i.e., numbers to subjective opinion) and “desubjectivization” (from SL opinions to another representation type), as demonstrated in Figure 28. Thus, in the process of subjectivization, a value of -1 was translated into disbelief, and a value of 1 was translated into belief. Consequently, in the process of desubjectivization, a value of the belief greater or equal to 0.5 was translated into a positive value (+1), a value of the disbelief greater or equal to 0.5 was translated into a negative value (-1), and a value of the uncertainty greater than the belief and the disbelief was translated into ambivalence (0). The opportunity to work with uncertain opinions is the greatest advantage of subjective logic for the purpose of this thesis, as well as the essence of the method for bias reduction presented in the next section.

To perform the subjectivization and desubjectivization, a MATLAB script was created for these operations, which enabled the analysis of the behaviour of the specific subjective logic operators for the different numerical values representing the trust score in the data sets. In that way, a method to interface subjective logic to existing formal apparatus is provided. This is crucial for subjective logic from a utility perspective, as it “pushes” its employment much closer to the hands of both the ordinary users and the skilled researchers.

In order to preserve a smoother flow of the text in this section, the concrete findings and numbers from the sensitivity analysis of the SL operators can be found in Appendix 9.2.3. However, here the implications from the results obtained are elaborated upon.

The results from the sensitivity analysis show that all subjective logic operators are very sensitive to small changes in opinion representation as a binary value substitute. For example, when binary 1 is represented as (0.8, 0.1, 0.1, 0.5), the performance of the operators differs significantly from the case where binary 1 is represented as (0.7, 0.2, 0.1, 0.5). The “desubjectivization” returns different results that would imply different performance compared to the previous case. The sensitivity (and the difference in

performance) is especially notable in the value of the uncertainty "*u*", which has the tendency to increase radically with small changes of the input. At first it may appear that this is a drawback of the operators, as it implies that they may not be robust to input changes and perturbations. However, when evaluating the operators on the data sets representing human relationships, it can be seen that it is actually in human nature to change opinions and trusting dispositions as a result of small changes in the behaviours of interacting entities. It is especially human to increase the level of doubt and uncertainty after a small and unexpected change. Hence, in a specific context and with the right subjectivization method, SL can be empirically shown to closely resemble human reasoning in the context of making trusting decisions.

On the other hand, the question that needs to be answered is: what is "the right" method of subjectivization and is it sensible to require one in the first place? The facilitating circumstance that enables this issue to be addressed is the fact that the subjective logic operator for opposite belief favouring should show identical performance as balance theory when applied to sentiment analysis (in the context of transitive trust reasoning). Therefore, the results obtained from the evaluation of the opposite belief favouring operator on the data sets are first compared to the results obtained from balance theory rules on the same data sets. The results show that for the values [0.8 0.1 0.1 0.5] for belief and [0.1 0.8 0.1 0.5] for disbelief we obtain identical performance for balance theory and the OBF operator. Clearly, the same holds if the belief is [0.9 0.1 0 0.5] and the disbelief is [0.1 0.9 0 0.5]. However, dogmatic beliefs (where $u=0$) are not used in the analysis, as they are unrealistic and they are not a subject of interest in this thesis.

5.2.4.2 Subjective Logic as a Predictive Model

Once the rules for mapping from binary value to subjective opinion and vice versa are established, the statistical evaluation of the subjective logic performance on the data sets can proceed.

Table 10: *SL operators' performance on the Slashdot data set.*

Slashdot	**Opposite Belief Favouring**	**Uncertainty Favouring**	**Base-rate Sensitivity**
Accuracy	91.35%	84.9%	84.9%
MCC	0.5654	0.1791	0.1791

Table 11: *SL operators' performance on the Epinions data set.*

Epinions	**Opposite Belief Favouring**	**Uncertainty Favouring**	**Base-rate Sensitivity**
Accuracy	93.18%	89.99% ~ 90%	89.99% ~ 90%
MCC	0.577	0.3471	0.3471

Table 12: *SL operators' performance on the Wikipedia data set (without accounting for the zeros).*

Wikipedia	**Opposite Belief Favouring**	**Uncertainty Favouring**	**Base-rate Sensitivity**
Accuracy	78.77%	73.46%	73.46%
MCC	0.2043	0.0906	0.0906

From the Slashdot data set, a total of 1,193,332 transitive triads were extracted. The results are summarized in Table 10. The best performance was exhibited by the opposite belief favouring operator, and it is the same as the balance theory performance. The other

two operators (uncertainty favouring and base-rate sensitivity) show identical performance to each other, and interestingly, relatively high accuracy but very low MCC. To explain why this is the case, the same discussion that was carried out for balance theory and its performance can be repeated here: the high accuracy falsely represents the model to perform well, while the MCC shows that it is actually unsatisfying. However, unlike balance theory, the subjective logic operators are also able to work with and predict neutral opinions. This is another reason for the seemingly lower predictive power of the two operators, in addition to the inherent positivity bias in the system. Furthermore, the high sensitivity of the UF and BS operators contributes to the big difference in the performance metrics compared to the OBF operator (and thus balance theory) as well. The Epinions data set (Table 11) is the largest, with 10,961,993 extracted transitive triplets. Interestingly, both of the performance metrics return the most satisfying results for this data set. This triggers the question: do the effects of positivity bias depend on the scale and the interconnectedness of the community members?

Table 13: *Wikipedia aggregated opinion-to-value representations and the SL operators' performance if neutral opinions are taken into account.* The subjective values for the sentiments are: - = [0.1.0.8 0.1 0.5], + = [0.8 0.1 0.1 0.5], 0 = [0.45 0.45 0.1 0.5].

	Opposite Belief Favouring	Uncertainty Favouring	Base-rate Sensitivity
(- o +)	[0.16 0.65 0.19 0.5] = -	[0.08 0.01 0.91 0.5] = 0	[0.12 0.015 0.865 0.5] = 0
(- o -)	[0.65 0.16 0.19 0.5] = +	[0.01 0.08 0.91 0.5] = 0	[0.015 0.12 0.865 0.5] = 0
(- o 0)	[0.405 0.405 0.19 0.5] = 0	[0.045 0.045 0.91 0.5] = 0	[0.0675 0.0675 0.865 0.5] = 0
(+ o -)	[0.16 0.65 0.19 0.5] = -	[0.08 0.64 0.28 0.5] = -	[0.085 0.68 0.235 0.5] = -
(+ o +)	[0.65 0.16 0.19 0.5] = +	[0.64 0.08 0.28 0.5] = +	[0.68 0.085 0.235 0.5] = +
(+ o 0)	[0.405 0.405 0.19 0.5] = 0	[0.36 0.36 0.28 0.5] = 0	[0.3825 0.3825 0.235 0.5] = 0
(0 o 0)	[0.405 0.405 0.19 0.5] = 0	[0.2025 0.2025 0.595 0.5] = 0	[0.225 0.225 0.55 0.5] = 0
(0 o -)	[0.405 0.405 0.19 0.5] = 0	[0.045 0.36 0.595 0.5] = 0	[0.05 0.4 0.55 0.5] = 0
(0 o +)	[0.405 0.405 0.19 0.5] = 0	[0.36 0.045 0.595 0.5] = 0	[0.4 0.05 0.55 0.5] = 0
Accuracy	75.66%	75.7%	75.7%
MCC	0.072	0.0814	0.0814

From the Wikipedia data set, 58,505 transitive triads were extracted (Tables 12 and 13). The same discussion on the performance of each of the SL operators as for the Slashdot data set applies here as well. The OBF exhibits the best performance, with the highest accuracy and MCC, although none of them offers acceptable performance despite the high accuracy. The MCCs of the UF and BS operators for the Wikipedia data set are quite low. This is probably due to the presence of additional evaluation choices (the neutral opinion). It is important here to note that the performance of the SL operators improves when accounting for the zeros in the data set (as shown in Table 13). This is also one of the advantages of subjective logic over balance theory, as well as over the rest of the computational models that account only for the binary inputs.

Since the UF and the BS operators are not commutative, there are two more opinions that should be taken into account: the (0 o -) and the (0 o +). The results from these operations are given in the last two rows, and they also make a difference in the predictions.

The high number of different subjective opinions for a neutral opinion makes it impractical to apply any of the UF and BS operators in a real system. Hence, the evaluation of the bias reduction method with these two operators makes no sense for the formal method. However, this does not make the statistical analysis less worthy, as the main interest is to evaluate the consistency between SL and BT. This is provided completely by the OBF operator. Hence, the bias reduction method is only applied to the results obtained from this operator.

Clearly, subjective logic offers a broader set of means for trust reasoning. Therefore, its performance on dynamic data sets is worth exploring in greater depth as well.

5.3 The Formal Model for Bias Reduction

The motivation behind the formal method for bias reduction is rooted in the fact that most of the theories of social interactions provide only a descriptive approach, and the analyses of the social phenomena that arise from those interactions are neglected. Thus, many of the properties that appear in larger networks cannot even be (accurately) described with the current models, let alone modelled or predicted using this approach. On the other hand, there is a trend of applying these theories as predictive models about virtual social networks evolving over time. This creates a possibility of system failure if some important properties have not been proven to hold. Another motivation for the bias reduction method is that the current trust systems and the computational algorithms for computing trust are created to mainly account for the designer aspect of the system or the algorithm. Thus, inherent in the design is the idea that each of the community members whose trust is to be calculated or who are to make the trusting decisions regarding some entity possesses the needed knowledge for such decisions and calculation. This is neither true nor can it be completely considered in human-centric systems, and it is essential that some mechanisms are developed that will compensate for such generalizations. By taking into account the available knowledge that the community has about the entity of interest, the method presented here answers this requirement.

In the remainder of this section, the formal background and analysis of the method for bias reduction are presented. Its implications for trust systems in general are discussed, and the third hypothesis of this thesis is resolved.

5.3.1 Extension of Transitive Trust Fusion

The bias reduction method is essentially an extension of the subjective logic operators for transitive trust fusion, in the sense that it integrates those operators and provides additional functionality. The employment of the three subjective logic operators is justified by the previous theoretical and statistical analysis about the link between transitivity and bias. The bias reduction method inherits all logical operators' properties and requires no formal proof of satisfying any special requirements and axioms from probabilistic logic. Despite being based on subjective logic, it adds several important features to the computational methods for trust reasoning that each of the SL operators separately does not possess. In essence, the method represents the discounting of a user's opinion obtained in an indirect way through transitive trust reasoning by the community's opinion about the user. In the formal language of subjective logic, this is written as:

$$\omega(A{:}\,C) = \omega^{C}_{community} \underline{\otimes}\, \omega(A \otimes B)$$

The operator $\otimes$ denotes some type of transitive fusion by applying one of the operators for uncertainty favouring, opposite belief favouring or base-rate sensitivity

(introduced in Chapter 2), whereas the operator $\underline{\otimes}$ denotes a dependent fusion between the community's opinion about C and the transitive trust chain between A and C (through B). Although the method appears to be simple, the notation hides a more complicated structure than it seems to represent. It is worth remembering here that integrated in each of the operators are numerous mathematical and logical operations. However, a complex mathematical representation does not need to be given here. The same logical notation is employed as in Chapter 2 and the bias reduction method is practically represented as a single operator embedding the properties of the operators it has integrated.

The issue of opinion dependence and its relation to the problem of user bias and trust computation was also introduced and discussed in the previous chapters (see section 5.2.3.3, Figure 27 a) and b)). Therefore, this is only briefly discussed here.

The simplest case of dependence is to assume that two agents (A and B) may not only hold identical information about an observed process or an entity, but that they also obtained it from a common information source. Furthermore, they believe that the observations they hold are independent, and treat (aggregate) each other's information as new and different. However, their opinions are clearly dependent, and a consensus on the information as if they were independent is meaningless to apply. It is important to note that in the framework of subjective logic, dependence is defined as having observed the same event (possibly at the same time) (Jøsang et al., 2006). However, in trust systems, the opinions to be aggregated will all be considered dependent if this is the definition of dependence. Namely, the opinion that each of the users gives about the entity of interest is essentially a result of observing the same entity. What is important to address in the context of trust systems is whether this opinion was obtained by an indirect encounter common for the two observers (or "trustors" in this case). Hence, it does not matter if the two observers make the same observations or if their estimates are equal. The main concern is to aggregate those opinions. The aim is to avoid repeatedly aggregating the opinions that originate from a common information source. The general consensus rule is based on the average of recorded positive and negative observations (see equation 2). The dependent fusion is also compliant with the general consensus rule in its goal and theory, with a slightly different formal definition that avoids the assumptions of statistical independence of the information sources. Its definition is as follows:

Dependent fusion (dependent consensus): Let us consider n agents A_i, where $i \in [1,n]$, and where $A_1, A_2 \ldots A_n$ hold some opinion about the proposition x. The dependent consensus is then

$$\omega_x^{A_1 \underline{\diamond} A_2 \underline{\diamond} \cdots \underline{\diamond} A_n} = (b_x^{A_1 \underline{\diamond} A_2 \underline{\diamond} \cdots \underline{\diamond} A_n}, d_x^{A_1 \underline{\diamond} A_2 \underline{\diamond} \cdots \underline{\diamond} A_n}, u_x^{A_1 \underline{\diamond} A_2 \underline{\diamond} \cdots \underline{\diamond} A_n}, a_x^{A_1 \underline{\diamond} A_2 \underline{\diamond} \cdots \underline{\diamond} A_n}),$$

such that:

$$\begin{cases} b_x^{A_1 \underline{\diamond} A_2 \underline{\diamond} \cdots \underline{\diamond} A_n} = \dfrac{\sum_1^n (b_x^{A_i} / u_x^{A_i})}{\sum_1^n (b_x^{A_i} / u_x^{A_i}) + \sum_1^n (d_x^{A_i} / u_x^{A_i}) + n} \\ d_x^{A_1 \underline{\diamond} A_2 \underline{\diamond} \cdots \underline{\diamond} A_n} = \dfrac{\sum_1^n (d_x^{A_i} / u_x^{A_i})}{\sum_1^n (b_x^{A_i} / u_x^{A_i}) + \sum_1^n (d_x^{A_i} / u_x^{A_i}) + n} \\ u_x^{A_1 \underline{\diamond} A_2 \underline{\diamond} \cdots \underline{\diamond} A_n} = \dfrac{n}{\sum_1^n (b_x^{A_i} / u_x^{A_i}) + \sum_1^n (d_x^{A_i} / u_x^{A_i}) + n} \\ a_x^{A_1 \underline{\diamond} A_2 \underline{\diamond} \cdots \underline{\diamond} A_n} = \dfrac{\sum_1^n a_x^{A_i}}{n} \end{cases} \quad (6)$$

where all the $u_x^{A_i}$ are different from zero. By using the symbol $\underline{\otimes}$ for this operation, we

can write $\omega_x^{A_1 \diamond A_2 \diamond \ldots \diamond A_n} = \omega_x^{A_1} \underline{\otimes} \omega_x^{A_2} \underline{\otimes} \ldots \underline{\otimes} \omega_x^{A_n}$.

The operator is both commutative and associative. The effect of the dependent consensus operator is to produce an opinion that is based on an average of positive and an average of negative evidence.

The discounting of a user's opinion by the general community's opinion about that user, and the utilization of the operator for dependent discounting, have several crucial implications for the computation of trust. First, it implies that, when assisting the trustor in making a trusting decision about whether to interact with a certain user (trustee), the method is concerned with the trustor's friends' recommendations about the trustee, and also with the trustee's overall reputation and his past performance. For the trust system in general, this provides a justification for the design assumption that the entities possess some knowledge about the rest of the network. This is important for the design of trust systems, and is currently considered to hold inherently. Another implication from the bias reduction model is that it accounts for the opinion dependence between the recommenders and the general community's opinion. This prevents inflation of ratings, which may happen if many dependent opinions are aggregated and considered independent. In that way, not only is the possibility of introducing bias in the calculation of the trust score taken into account, but the method deals with information dependence from the beginning of the trust aggregation.

It is important to note here that statistical analyses with the operator for independent fusion (the general consensus operator) were also performed. The application of this operator resulted in unrealistically high beliefs, even higher than the beliefs that each of the entities holds about the entity whose trust is to be assessed. Therefore, its use for the purpose of countering user bias was overruled.

5.3.2 Evaluation and Analysis

In this section, we present the results from the statistical analysis of applying the bias reduction method to data obtained from real-world implementations of the trust systems of Slashdot and Wikipedia. The statistics regarding the number of nodes, edges and extracted transitive triples in the data sets are the same as given in the previous section. Here, we only elaborate on the results obtained from applying the method to the data sets. They are given in Table 14 and Table 15, for the Slashdot and the Wikipedia data set, respectively.

Table 14: *SL operators' performance on the Slashdot data set after applying the bias reduction method.*

Slashdot	**Opposite Belief Favouring**	**Uncertainty Favouring**	**Base-rate Sensitivity**
Accuracy	90.25%	89.8%	89.84%
MCC	0.3433	0.2989	0.3017

Table 15: *SL operators' performance on the Wikipedia data set after applying the bias reduction method.*

Wikipedia	**Opposite Belief Favouring**	**Uncertainty Favouring**	**Base-rate Sensitivity**
Accuracy	78.78%	/	/
MCC	0.1564	/	/

By comparing the results of the calculations obtained from applying the uncertainty favouring and the base-rate sensitivity operator without the bias reduction method, it can

be concluded that the uncertainty favouring and the base-rate sensitivity operators relax the trusting decision of a transaction by lowering the uncertainty in the decision-makers and helping them to come to a more informed decision. More importantly, this affects the overall propensity for interaction among the community members, enabling the virtual analogue of trust to come much closer to the traditional concept of trust – a "social glue".

In the Wikipedia data set, it was found that if the number of overall positive ratings was compared before and after applying the bias reduction method, the obtained result was slightly higher in the raw data set (49,332) than in the data obtained by applying the method (49,004). This result is to some extent expected (although it is not the rule), since the natural tendency of users is to give positive feedback. Hence, biased results are inherently present in the data. This effect of increased positives in the Wikipedia data set and decreased in the Slashdot data set may be used as an argument about the benefits of applying shades of grey (such as the neutral opinion in Wikipedia) together with the application of the bias reduction method. The bias reduction method affects the rating distribution in a way that leads to reducing the bias. Applied in a real-time setting, it is also expected to limit the amount of possible bias that can be exhibited by users. Nevertheless, it is important to realize that the application of this or other formal methods for bias reduction is only one part of the whole framework aimed at addressing the problem of user bias in online trust systems. It is essential to embed the whole set of measures recommended by the previous analysis and empirical studies in order to build an efficient and effective platform for online trusting interactions.

Figure 29 compares the performance of balance theory, subjective logic and the bias reduction method through the receiver operator characteristic, known as "the ROC curve". The ROC curve is a graphical plot that illustrates the performance of a binary classifier system while varying its discrimination threshold. The curve is obtained by plotting the true-positive rate (TPR) against the false-positive rate (FPR) at various threshold settings. (The true-positive rate is also known as sensitivity, or recall in machine learning, and the false-positive rate is known as "the fallout", and can be calculated as 1-*specificity*.) The ROC curve is, thus, the sensitivity as a function of the fallout. The best possible prediction method would produce a point in the upper left corner or coordinate (0, 1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). This point is also called a *perfect classification*. A completely random guess would yield a point somewhere along the diagonal line (the so-called *line of no discrimination*) from the bottom left to the top right corner (regardless of the positive and negative base rates).

From Figure 29, it is evident that balance theory shows better performance than the SL-based bias reduction method: the whole BT ROC curve (represented by the blue line) is higher from both the line of no discrimination and above the SL OBF ROC curve after applying the bias reduction method (represented by the red line). However, this appears to be the case only because the results are evaluated on a data set (in this case the Slashdot data set) that includes a significant amount of positivity bias and does not account for ambivalent opinions. The bias reduction method, while still offering a decent amount of fitness to the results, enables the application of all three SL operators for transitive trust reasoning and accounts for ambivalence. This is crucial for further contextual analysis and modelling of the entities' trusting dispositions in their online interactions.

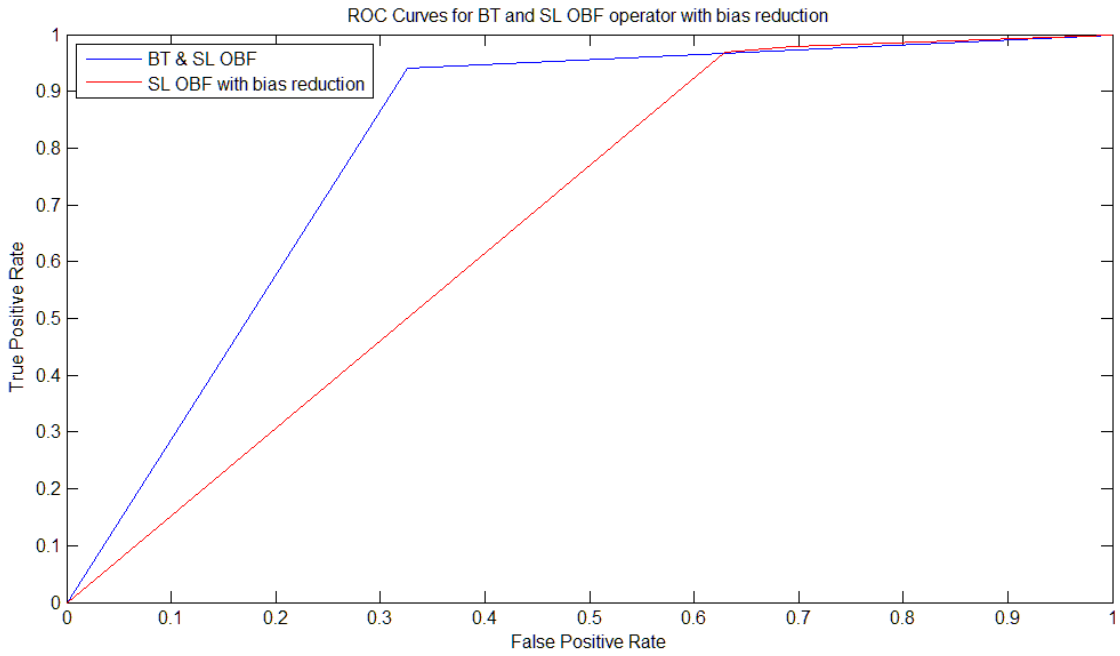


Figure 29: *Comparative analysis of the performance of balance theory, subjective logic and the bias reduction method through the ROC curves.*

A major contribution from applying the method is the possibility of a seamless application of the methodology for trust reasoning offered by subjective logic. In this regard, it is an improvement over the ordinary subjective logic transitive trust reasoning. Moreover, it is well integrated in the framework for alleviating user bias proposed as a result of the empirical study in the previous chapter, as it enables reasoning with a broader spectrum of opinions (such as the ambivalent opinion in the Wikipedia data set) and other "shades of grey" types of user feedback. Finally, as it accounts for the subjectivity of user reasoning, it is more adaptable for picking on and embedding the hidden signals from the user behaviour, which is essential for the soft methods offering decision support. Hence, the bias reduction method offers greater and better expressiveness in the computational methods for trust reasoning, regardless of the way the trust score is represented (i.e. regardless of the requirement for input for the trust score in the computational engine).

For the general design principles that this thesis is proposing, it is crucial to realize where in the taxonomy tree of the systemic framework (introduced in Chapter 3) the bias reduction method fits. Given the D-E-C nature of trust systems, by which they are categorized as dynamic, environment-oriented and densely connected systems, it is clear that accounting for subjectivity, transitivity, bias and the history of entities' behaviours makes this approach a great contributor to all three categories. However, it is mainly the connectedness (the C property) and the purposiveness (the environment orientation and the E property) that it aims to address, and it is less concerned with the dynamicity of the interactions and the D property of trust systems. These two were also identified as the characteristics that are most rarely tackled simultaneously by the same approach. Delving more deeply into the taxonomy factors, the bias reduction method affects and is greatly affected by the active entity behaviour (AEB), the context, and the resilience and evolutionism, and partially by the time and dynamics factor. Hence, it is incorporated (as a cause and/or effect) in the entire taxonomy for designing human-centric systems.

5.3.2.1 Knowledge of the Network and the Observer's Problem

Unlike balance theory, which assumes that each of the users possesses some knowledge about the rest of the network, subjective logic does not make such assumptions. Even when such knowledge is not available, subjective logic is able to derive an opinion based on any knowledge available from the overall network connections about the interactions with the user of interest. However, there is currently no model offered to employ and integrate these advantages for practical purposes. This is the foundation of the formal method for bias reduction: using both the general standing of the community (the network) in the form of an *overall reputation* of a certain user and the knowledge that a closer circle of the user's neighbours has in the form of *transitive trust reasoning*.

Not taking the viewpoint of each of the entities in the triad directly affects the prior probability distribution of trust opinions (and with that, the base rate included in the subjective opinion) by disregarding the number of entities whose opinions were considered in order to derive the trust score/rating. This has a direct impact on the metrics for transitivity and triadic census, but also on any other metric for social network analysis. Because base rates and conditionals are not normally considered in traditional Dempster-Shafer belief reasoning, there is a risk of falling victim to the base-rate fallacy in practical applications (Josang and O'Hara, 2010). The bias reduction method is able to incorporate the base-rate sensitivity operator in its trust reasoning algorithm, accounting for the danger of this type of probabilistic fallacy.

5.3.2.2 Opinion Dependence and Conflict

In this section, the issues of dependence and conflict are analysed. These were also discussed for balance theory, with the difference that in the following analysis the problem of using subjective logic and applying the bias reduction method is dealt with.

In order to be certain that balance theory holds, let us assume that all of the directed edges are positive, and let us assign the value for all the opinions as $w = [b, d, u, a] = [0.8, 0.1, 0.1, 0.5]$, according to the results obtained from the sensitivity analysis. By "all opinions", we refer to the initial (non-aggregated) opinions $w_D^A,\ w_D^B,\ w_D^C,\ w_X^D, w_A^E, w_B^E, w_C^E$, represented by non-dotted arrows in Figure 30 a).

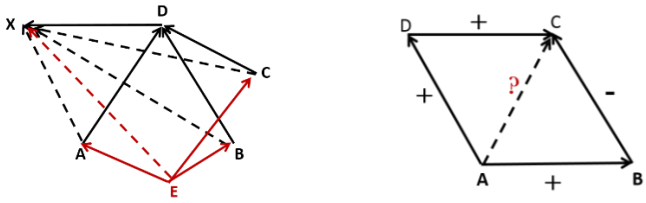


Figure 30: *a) Opinion dependence* *b) opinion conflict.*

By only employing the consensus operator to aggregate the available opinions, an important part of the knowledge in the network represented by the direct interactions among the community members will be disregarded. Moreover, this will result in overly positive opinions, whose belief is even higher than the initial beliefs of the opinion sources.

For instance, by computing the opinion of A about the value of proposition X, the following options are provided:

$$w_X^A = w_D^A * w_X^D \text{ or } w_X^A = w_B^A * w_X^B = w_B^A * w_D^B * w_X^D$$

where $*$ denotes some operation with subjective logic. The two possibilities for calculating A's opinion about X may produce different results. This depends on the chain

of trust that one chooses to go along, and on the SL operations between the opinions (for example, if opinion dependence is considered or simply an independent opinion aggregation is performed). Different results can also be obtained if the calculation is done by accounting for the transitive triplets rather than going along a direct one-hop or two-hop trust chain. However, if only transitive trust aggregation is applied, reasoning about very near neighbourhood can only be possible. Moreover, the issue of bias will not be tackled in any way. Finally, simply applying the dependent fusion operator, a significant part of the knowledge about the network (the direct opinions of the recommenders about the interaction entities) will not be used.

If transitivity is disregarded, independent fusion gives $w_X^A = [0.71, 0.1, 0.19\ 0.5]$, whereas dependent fusion produces $w_X^A = [0.59, 0.07, 0.34, 0.5]$. If w_X^A is computed by accounting for the transitive relations as well, the independent fusion will give $w_X^A = [0.68\ 0.19\ 0.13\ 0.5]$, and the independent $w_X^A = [0.6\ 0.17\ 0.23\ 0.5]$. It can be seen that the dependent consensus produces opinions with lower belief and higher uncertainty. While this prevents inflation of positive ratings, it may lead to situations where the decision-maker cannot come to a decision because of the increased level of uncertainty. The bias reduction method accounts for this by distributing the uncertainty to the belief mass (i.e. to the belief and the disbelief), while increasing the level of information the decision-maker possesses by providing him with the community's opinion about the trustee.

Remember, balance theory does not offer any means for dealing with opinion dependence and opinion conflict. Moreover, it implicitly considers users' sentiments to be independent of one another in the formal representation. At the same time, it considers their friending affinity to be highly dependent on knowledge about the network. This is not only a constraining condition for the practical analysis of real-world data, but it is also a contradictory assumption for the design of trust systems.

The application of the bias reduction method in the analysis of small-scale trust systems offers a small pay-off, as in a small and closed community it is easier to assess someone's trustworthiness from close encounters rather than through heavy logical reasoning. However, it offers theoretical support for the proof of concept through the given examples.

When it comes to the ordinary consensus operator that deals with conflict resolution, one contradictory requirement that is presented for the computational engine is the need for opinion and information source independence. Although this requirement renders it useless for bias reduction, it may be necessary for some contexts to assume independence for practical reasons. Therefore, we leave it to future work to investigate the possibility of including it in a single context-dependent operator with an integrated decision mechanism for context-switching trusting interactions.

For clarity, the formal model for alleviating user bias in trust systems is presented in this chapter in a more condensed procedural form:

1. Data preprocessing // unification of the data sets' representation
   represent data in the format: (source node, destination node, sentiment)
   extract transitive triads: apply Algorithm 1
   extract triplets' signs
2. SL operators programming
   conflictFusion()
   dependentConsensus()
   discounting()

baserateSensitivity()
oppositeBeliefFavouring()
uncertaintyFavouring()
biasReduction()

3. Subjectivization
   perform sensitivity tests //determine the opinion-value pair for the given context
   convert values (sentiments) to subjective opinions
   apply to data set (see Figure 28)
   apply balance theory
   apply SL operators of interest
   apply.biasReduction()
3. Desubjectivization
   convert opinions to numerical value
4. Evaluation
   compare results (signs) to the original data sets
   apply performance metrics (accuracy, MCC, ROC curves)

It is important to note that this procedure is not to be taken as a step-by-step guide to developing a model for bias reduction. If such a model is needed, the concrete design will depend on many factors, such as the context, and the type of biases present in the system.

5.3.3 Extensions of the Bias Reduction Method

The bias reduction method presented in this thesis provides a foundation for a reasoning model development that can be transferred to a general methodology, with the possibility of further extensions.

5.3.3.1 Applying Evolutionary Principles

One possibility for extending the current model for bias reduction is through exploitation of some principles of evolutionary theory and by applying randomness to the methods for information aggregation. Evolutionary psychologists argue that much of human behaviour is based on the output of psychological adaptations that evolved to solve recurrent problems in human ancestral environments (Confer et al., 2010) . The theory and its ideas are also compliant with the ideas promoted by (general) systems theory, especially when applied to the field of evolutionary computation (Cagnoni, 2000).

Applied to the bias reduction method, it could take the following form:

$$\omega(A:C) = (\omega^{C}_{Community} \otimes \omega^{B}_{A}) \oplus (\omega^{Subset(A'sNeighbours)}_{Community} \otimes \omega^{B}_{Subset(A'sNeighbours)}) \oplus (\omega^{Random(B'sNeighbours)}_{Community} \otimes \omega^{B}_{Random(B'sNeighbours)})$$

The ultimate goal of applying elements from evolutionary theory to the bias reduction model developed here would be to increase the intelligence of the algorithm for choosing information sources. This will result in increased diversity of opinions. To enable this extension, the idea of exploration and exploitation was used from evolutionary computation. An approach that employs similar ideas is the BarterCast reputation mechanism (Meulpolder, Pouwelse, Epema and Sips, 2009b).

5.3.3.2 Trust Transitivity and Conditional Belief Reasoning

Given the nature of human trust transitivity, any simple mathematical model will essentially have poor predictive power. Therefore, a new interpretation of trust transitivity is needed when it comes to a virtual setting. For example, instead of simple transitive trust reasoning, recommendations from an advisor can be considered as the input arguments to a conditional reasoning model for assessing hypotheses about the trustee. Such a model of conditional trust transitivity has already been considered as a possibility (Jøsang et al., 2012). The idea here is to model trust transitivity in a similar manner to analysing competing hypotheses, as applied in the area of intelligence analysis (Pope, 2005). Modelling trust transitivity as a simple evidence analysis problem removes the mysticism from trust transitivity, as it does not assume that it exists as a separate natural process. In addition, the model accounts for weighing of the obtained evidence in order to draw conclusions about the trust target. This enables the principle of trust propagation to be general and flexible, but moreover applicable to the online communities of people, organizations and software agents. The higher purpose, however, would be providing decision support based on collaborative interpretation of evidence by the community members. Consequently, this leads to enhancing the ability of online communities to support the emergent properties that result from the transitive trust relations of their members.

While the previous three interpretations of trust transitivity can provide intuitive results in different situations, it is still problematic that the analyst needs to select one of them to model and analyse any specific situation. This calls for automated decision algorithms that will allow the application of subjective logic to larger-scale scenarios, and with that, to practical implementations as well.

5.4 Conclusion and Discussion

This chapter provided statistical analysis (for the larger part) and theoretical analysis of both balance theory and subjective logic, as well as the newly proposed bias reduction method, and their performance in relation to real-world data sets. A major goal was to investigate the utility of subjective logic as a formal apparatus for trust reasoning in a real-world setting. To do that, we have performed tests on each of the operators that have been developed in this framework. The results from these tests are given in the Appendix at the end of the thesis.

We have analysed balance theory and subjective logic regarding their predictive power for relations formation and established a connection between the two through the subjective logic operator for opposite belief favouring. This approach was verified with the statistical analysis of real data. The approach offered the opportunity to use balance theory as a baseline for deciding about the correct method for subjectivization/desubjectivization. On the other hand, in order for such a mapping to hold, balance theory had to be proven to perform well as a reasoning method in human-to-human relationships. Although balance theory had already been proved to be a good representative of social interactions by other researchers, we repeated the evaluations for the concrete data sets in order to be able to conduct a broader analysis. The results on the performance of balance theory turned out to be satisfying, but more careful analysis of the causes and effects of the results obtained showed that this happened to be the case only because of the inherent inclination towards positivity bias in the user feedback. Similar analyses were carried out for each of the subjective logic operators for transitive trust reasoning.

This study is the first to practically evaluate the performance of subjective logic and

offered important insights into the utility of this logical framework. Despite the significance of the findings in the presented research, there are still questions that are left unanswered. Some questions are related to the applicability of subjective logic to the broader context of trust systems design and operation. Questions also arose about the context of the transitive trust reasoning. We believe that none of these questions have a simple answer. For instance, the mapping from normal values to subjective opinions is very context-dependent. Moreover, there is also ultimately no "right" way to represent a trust relationship with normal integer values. Therefore, while we can evaluate the interchangeability between normal values and subjective opinions, this still does not mean that there is a unique or even correct representation of trusting relationships.

One obvious advantage of subjective logic is that it also enables reasoning with the intensities of the user opinions (how positive, how negative). Unlike balance theory, subjective logic operators are also able to express and work with ambivalent opinions. Furthermore, subjective logic takes into account the number of hops after which translating belief from one entity to another makes sense. Balance theory does not employ such mechanisms as it does not account for situations that involve recommendations over more than one hop.

In addition to the general advantages of subjective logic, one peculiarity of SL operators regarding transitive trust fusion is that they take into consideration the order in which the opinions are combined. This is a very important property of subjective logic because it promises additional methods for trust calculation. All of the operators assume independence of opinions when applied to trust chains longer than one hop. The bias reduction method developed in this work "corrects" this properly, because it accounts for the general community's opinion in addition to the use of transitive fusion to one-hop relationships.

On the other hand, it is a disadvantage of subjective logic that it takes distrust/disbelief to be equal to absence of trust. The uncertainty parameter is responsible for this to some extent. However, the question arises here: what happens with non-observed events? And what happens when the same entity combines different versions of the same proposition? Subjective logic considers the absence of evidence as evidence of absence. This is considered here to be an issue of context dependence in the transitive trust operators. This enables exploitation of the scenarios where context similarity offers interoperable formal apparatus for calculating trust.

The bias reduction model proposed in the thesis tries to address the issue of dynamics by accounting for the general community's opinion about the entity of interest rather than employing special forms of heavy logical aggregation. As the formal method is also based on subjective logic, its employment is justified by the previously performed statistical analysis. Its performance is evaluated on data sets from real-world trust systems and discussion is offered on the advantages and the possibilities for improvement and extension of the method. Such analyses are essential, as subjective logic has mainly been based on theoretical and analytical results prior to this work. The SL was intended to be used mainly for theoretical and analytical purposes, and there is little proof that demonstrates and supports its practical significance.

A persistent problem exists with the task of aggregation of dependent opinions, as such opinions have to be detected in the first place. With the presented method, it is possible to disregard this requirement and apply and verify the theories for aggregation of dependent opinions on isolated examples. However, it still seems impossible to find the amount of dependence between the offline and the online connections. This has an impact on trust establishment online, and has not been considered so far by any of the currently proposed models. Regardless of the number of such connections (even regardless of their strength), their impact may be the most significant determinant in the trust system evolution.

Accounting for opinion and source dependence is also crucial for enabling the exploitation of the so-called *wisdom of crowds* (Surowiecki, 2005). This requires the availability of diverse opinions that are not determined by the rest of the community's opinions, and that a large set of such opinions exists on which to base the general opinion. Surowiecki himself noted that, even in the traditional societies, homogeneity, centralization, imitation and emotionality are the main factors that lead to information cascades, herd instincts and cases of mass hysteria – all of which are manifestations of people's cognitive biases. This undoubtedly holds for large-scale virtual societies, where the effects of these phenomena are felt much faster and have a greater impact on the system.

Currently, there is no tool that implements subjective logic as a reasoning engine, as is the case with fuzzy logic and probabilistic logic. This can be considered a downside of subjective logic, as having such a tool would increase significantly its utility and pragmatism. Moreover, it would enable subjective logic to be interfaced with other paradigms that already have their utility interface, directly proving its backward compatibility to the existing logical apparatus. Although this has already been discussed theoretically and analysed formally, as a natural way of human reasoning, we believe that subjective logic can reach a much greater level of exploitation not only by researchers, but also by ordinary users.

A major contribution of this chapter, as well as the thesis in general, is the practical implementation and analysis of the performance of subjective logic apparatus on real-world data. This also provides novel material for theoretical analysis of subjective logic operators and methods. However, to justify the problem statements and the ideas presented in the thesis, a broad range of techniques and tools from all the social interaction theories introduced in Chapter 2 was employed. Transitivity and triadic closure in the real-world data sets were analysed with Pajek, a tool designed for analysis and visualization of large networks. To study the context similarity between the data sets and the analysed trust systems, the network motifs were extracted. This enabled the establishment of the three data sets in the appropriate superfamilies of evolved networks with known properties. In addition, the statistical significance of each triad type was confirmed and further comparative analyses were carried out between the extracted profiles of the three data sets. Thus, everything that had been discussed theoretically in the previous chapters was demonstrated practically in the final chapter. With this, the way towards meaningful statistical analysis of the performance of balance theory and subjective logic was open, allowing the design of a framework for bias reduction based on formal, technical and social mechanisms.

6 Concluding Remarks

6.1 Summary of Contributions and Results

In the end, it useful to return to the main hypothesis, which is also the main contribution of the thesis: adequately structured aggregation of the trust information gathered from sources with diverse and dependent opinions can help in alleviating and limiting the effect of user bias on the trust system. Hence, the thesis as a whole represents a methodology for designing trust systems with limited amount of bias. It is the first work in the field of computational trust to tackle the issues of user bias in an empirical, statistical, and formal manner. One of the major contributions is applying multidisciplinary approaches in the resolution of computational and technical issues. In that context, the thesis determined the systemic properties of trust systems that are lacking in the current technical solutions and incorporated them in a novel taxonomy of trust systems, which serves as guidelines for designing online trust systems.

Another contribution of this work is the performed empirical study of reputation, recommendation and rating systems demonstrating empirically the presence of biased behaviour of the users. For that purpose, it identified and studied for the first time in the trust research several types of concrete cognitive biases that the users of trust systems exhibit in their online experiences. This allowed for the design of a framework for identifying and tackling the bias phenomena for the purpose of providing good user experience, system reliability and result accuracy. The significance of the methodology presented in the thesis is in the revelation of trust as a core aspect in the solutions for increased complexity and user bias.

The main contribution, however, is the design and proposal of models and mechanisms for aggregation of trust information in online trust systems under biased user behaviour. Through statistical analysis of large data sets from real online systems, the applicability of the proposed methods was justified and empirically evaluated. Moreover, new insights into the management of trust information in the presence of dependent information sources were provided and, as a consequence, the state of the art was advanced in both the field of aggregation of trust information and the field of network (community) structure.

In the process of investigating and designing the means for identifying and limiting the effects of user bias on online trust systems, the thesis studied the application of techniques and methods from the fields of information fusion, behavioural science, and social computing. Based on that, it proposed adequate solutions for the problems of trust systems' design and trust aggregation under user bias.

More precisely, it applied general systems theory and case study research for the purpose of theory building, taxonomical categorization and evaluation of technical human-centric solutions. In that context, a novel taxonomy of trust systems was introduced for categorization of the subcomponents and the design principles of trust systems. With that, the thesis demonstrated the lack of important elements (from a systemic perspective) in the design of human-centric trust systems. By applying prospect theory and choice architecture, a framework is designed for empirical identification and detection of user bias in a trust system. Finally, methods from network theory and subjective logic were employed in the design of a formal method for limiting the amount

of bias in trust systems. The accounting for dependence, transitivity and subjectivity as contextual parameters of the mechanisms for trust aggregation was achieved by employing information fusion, with some necessary improvements. For that purpose, the dependence between the user bias and the information representing the social relations in the trust system was established. This allowed the effects of bias to be quantified and affected in a desirable manner in a methodology that accounted for the transitivity, subjectivity and dependence as properties of the gathered information.

With that, it was demonstrated that adequately structured aggregation of the trust information gathered from sources with diverse and dependent opinions can help in alleviating and limiting the effect of user bias on the trust system, i.e., the main hypothesis was confirmed.

6.2 Conclusion

Trust is a social phenomenon, which must be defined for a particular context. It can be both a feeling and a model, an idea and a reality. Regardless of the context, trust has people at its core. Understanding how all of its representations work and how closely they resemble each other is essential for the design practices of trust systems, regardless of whether they represent the social interactions in the traditional or in the virtual world. Our work detected and analysed context, risk, dynamics and evolutionism, among many others, as factors that largely influence both the work of the trust systems and the users' understanding of that work.

Earning online reputation and gaining the trust of the community, although a social process, differ significantly depending on the nature of the system that integrates the interactions. Nevertheless, the success of both the online and the offline environments depends largely on the trusting choices that the system entities make during their interactions. It has been known for the property of trust to act as a social lubricant in the interactions between people. The possibility of realizing numerous transactions and virtual experiences in a short time requires scalable solutions for establishing trust and reputation. In such a dynamic environment, having no social mechanism to adequately capture the system evolution is equal to being well equipped for a world that no longer exists.

Incorporating the social elements into the technical operation of virtual platforms, and accounting for the organismic traits of the entities' interactions, lead to the emergence of complex social phenomena. These phenomena are not necessarily the product of the rules that guide the behaviour of the individuals themselves. The high discrepancy between the designers' and the users' expectations and the trust system performance can be seen in the persistence of malevolent and biased behaviour. The manifestation of such behaviours is seen in the scores and the ratings obtained from trust systems, but also in the quality of the ideas and the products met in the virtual world. However, such phenomena are not necessarily the result of erroneous or malicious inputs by individuals. It is important to consider the whole chain of entities' dependencies in the trust system in order to prevent the reinforcement of undesired phenomena.

A significant mark of the thesis is that it discusses important issues that stem from the human-centric nature of computational trust systems, such as: the issue of the "entities as observers" in the triad, the entities' knowledge about the network, the significance of neutral opinion as a choice in the trust system and its effect on positivity bias. The main aim of the thesis is to serve as a testimony for the general research community that accounting for the social phenomena in human-centric computational systems is a crucial part of both the design and the maintenance of those systems.

In the introductory chapters, a theoretical foundation was provided for the topics presented in the thesis, and the problem statement was formulated. The later chapters were directed towards confirmation of the theoretical analysis and the provision of sound formal support thorough statistical and empirical analysis.

Chapter 1 outlined the problem statement and the research goals of the thesis, pointing out the major hypothesis and the contributions of this study, but also describing the methodology employed to achieve these goals.

Chapter 2 presented several theories of a social and computational nature that are important for understanding the issues and the hypotheses resolved within the thesis. The significance of each theory for the particular issues was specified and a broad description of related work was provided in order to fit this work into the state of the art. Special attention was devoted to balance theory and subjective logic as the main theories employed and analysed throughout the thesis. The formal apparatus that each of them offers in the exploration of social interactions was also presented in this chapter, together with some theoretical analysis of the central issues.

Chapter 3 presented a taxonomical framework as a systemic approach for designing dynamic, densely connected and environment-oriented trust systems. It proposed improvements to the existing work by defining four design factors and determining their relation to each of the trust system components. It was devoted to resolving the first hypothesis from the problem statement. *Active entity behaviour, context, resilience and evolutionism*, and *time and system dynamics* were introduced as major influencing factors on the trust system design. These insights were incorporated into a multidimensional taxonomy together with the trust subsystems in order to establish the interconnections and dependencies between them. Moreover, the users were placed at the core of the trust systems. The framework was represented as a series of taxonomical trees, and a granular categorization of the trust system properties was introduced. Thus, a broader representation of the design choices was offered. Moreover, a notion was defined of the emergent properties of trust systems and the points where these properties may work against the designer's intuition were located. By performing semantic analysis of the state of the art of trust systems, two representative case studies were chosen for evaluation upon the framework. This work is the first to apply fundamental social science principles, more particularly, principles from general systems theory and case study research for the purpose of theory building and evaluation of technical human-centric solutions. By doing this, it joins the creative vigour of soft systemic analysis with the quantifiable practicality of "hard" technical solutions. Finally, by mapping BarterCast and Yahoo! Answers as representative case studies of socially inspired trust systems onto the framework, important weaknesses were revealed and some improvements were proposed. This chapter offered a blend of social and computational principles aimed at responding to the challenge of introducing systems thinking into the research on computational trust.

Chapter 4 analysed user bias as one of the greatest causes of malfunctioning of the current trust systems. It addressed the second and the third hypothesis of the problem statement. The relation between computational trust systems and user behaviour was thoroughly explored. Several types of concrete biases that are exhibited by users in their online experiences with trust systems were analysed. Many additional interdependencies were revealed among the social phenomena present in trust systems that have not been captured by the current models. In an empirical study that included nearly a hundred users, it was found that the users exhibit distinction bias and positivity bias in both phases of decision-making: anchoring and framing. Based on the results obtained from the empirical study, a choice architecture for trust systems was proposed to deal with the issues of biased behaviour. It addressed the possibility of employing hidden signals in trust systems and providing a greater granularity of the choice attributes and options in

order to minimize or alleviate user biases.

Chapter 5 was devoted to formal modelling and statistical analysis of the methods for bias alleviation. In addition, theoretical analysis of balance theory and subjective logic was provided and their performance on three real-world data sets was evaluated. This chapter accounted for the load of the hypotheses in the thesis. The goal was to investigate the applicability of subjective logic as a formal apparatus for trust reasoning in a practical setting. The utility of subjective logic was thus improved, as its functionality was practically demonstrated through an enhanced model for bias reduction. Each of the subjective logic operators for transitive trust reasoning was also analysed, discussed and considered for improvement before inclusion in the proposed model for bias reduction. Despite giving a major proof of concept for the envisaged hypotheses, this chapter wrapped up the ideas from the new taxonomy, making the connection between the contributions of the thesis and the D-E-C nature of trust systems defined in Chapter 3.

It is important to note that the meaning of bias reduction or *alleviating user bias* is not annulation of bias in trust systems, but limiting the effects of biased user behaviour on the functionality of online trust systems. It is one thing to talk about user behaviour, and another to talk about the effects of a certain behaviour on the system. It would be unreasonable to state control over the former, as the baseline of a correct user behaviour is not determined by the system designer. A large body of research has established that a defining characteristic of cognitive biases is that they manifest automatically and unconsciously over a wide range of human reasoning. This is one of the reasons why most of the proposed approaches for bias mitigation are not formalized. Rather, soft non-technical solutions are proposed that have the aim and potential to increase awareness of the possibility of cognitive bias and the options for dealing with such contexts.

Clearly, it is extremely hard to evaluate any methodology that is aimed at revealing or treating the human-centric social phenomena in the virtual world. Cognitive bias mitigation is the prevention and reduction of the negative effects of cognitive biases. However, there is no coherent, comprehensive theory or practice of cognitive bias mitigation – only tools, methods, proposals and other initiatives – in academic and professional disciplines concerned with the efficacy of human reasoning. Most of them address mitigation tacitly rather than explicitly. It is also unrealistic to base the expectations from these phenomena on the ones met in the real world, as the two are not interchangeable with respect to human behaviour. The results from the statistical analysis trigger the question of whether the term “bias” is correct when it comes to describing what is often reasonable human behaviour of “acting like the group”, “leaving feedback only out of concern”, etc. Reasoning in a holistic manner is one of the main strengths of this work.

We (the research community) tend to confuse the difference between generalizing concepts to create common vocabularies in the field and generalizing the rules and principles of research in the field. Whereas the former is essential for defining and establishing the scientific background, the latter is highly context-dependent and does not offer the same ease of creating universal laws. In other words, the former facilitates the exchange of scientific ideas and results, but the latter hinders the progress of the field by practising reductionism. It is important to recognize this difference and to understand how to comply with the needs and requirements for designing and tackling human-centric systems.

6.3 Future Work

Trusting behaviour is an emergent phenomenon that evolves from the dynamic interactions between socio-economic, technical and biophysical entities in a

computational system. In addition to providing guidelines for systemic reasoning about trust, this work is aimed at raising awareness that the trust system as a whole should not be analysed only through the behaviour of its parts. Considering the inter-connections between those parts is important. This aim will also lead any future research related to human-centric systems. The framework introduced in this thesis will also evolve with any additional insights obtained from the research. It will be guided by the principles of general systems theory and fitted in the current taxonomical tree organization. Additional factors influencing the design of human-centric trust systems are also expected to be defined and to become part of the multidimensional map of the interdependencies between the system entities.

A major part of the future work will be the research on community structure in trust networks in order to account for the quality of relationships between the system entities (positive/negative – sentiment polarity, as well as sentiment intensity). These insights will be employed to study the convergence of these relationship types in various social contexts in order to model the diffusion of trust opinions and beliefs in a more dynamic manner. Such an approach would help the evaluation of the work on other social interaction theories except those elaborated in this work.

An inherent disadvantage of subjective logic is that it does not provide means to qualify and quantify the presence of source dependence in the trust system. The practicality of most of the logical frameworks is shadowed by their strong analytical constitution, making them hardly applicable to technical systems. Therefore, means for demonstrating the utility of the formal model are required. This translates to a requirement to design a separate utility tool that will allow the conditions under which dependency and network structures emerge to be studied, and their effect on user bias analysed. This model should then be interfaced with the existing formal apparatus and its performance and efficiency in relation to real-world data should be evaluated.

Future work will also concentrate on investigating in greater depth dynamics as a design factor. Perceptual control theory was referenced as part of the formal computational apparatus that complies with the systemic principles proposed in the thesis, which tackles the issue of dynamics from an engineering perspective as well. Together with subjective logic as a framework that provides formal reasoning with human beliefs, the main part of future work will be directed towards joining the ideas of these two theories to formalize the evolving trust relationships under biased user behaviour.

Featuring the dependence of information sources and the diversity that can arise from the “competing” opinions of the users in a separate model is another issue that has not been explored by the current research on computational trust. There is no method for quantifying the effect of user bias on trust systems, especially with reference to unbiased systems. Although the current approaches evaluate their models’ performance by comparing them to other models in the field, knowing the reference for non-biased results is important for determining the trade-offs between cost and complexity, usability and robustness. Evolutionary theory has already been recognized by many computational disciplines as a source of ideas related to algorithm design and performance optimization. The concepts of exploration and exploitation have been directly related to the property of adaptability of a system through the diversity of system entities. However, in the research on trust systems, which are essentially human-centric and dynamic, diversity has neither been explored nor connected to the system’s adaptability. Thus, we add these evolutionary principles to our future work.

7 Acknowledgements

I would like to thank my supervisor, Borka Jerman Blažič, for her understanding, support and guidance. She hides a source of never-ending enthusiasm that I have found amusing to rediscover. She always remained available regardless of her many duties and projects. Her insights into my work brought about the most significant improvements in the thesis.

I am also grateful to have had Tomaz Klobučar and Dušan Gabrijelčič as co-supervisors. They are not only the most devoted workers I have ever met, but also great friends and colleagues. They are a true example of how perfectionism differs from niggliness.

My experience at Jozef Stefan Institute and the Laboratory for Open Systems and Networks has been the most enlightening period of my life. I am most thankful to all my colleagues that made it that way, most of all to Tatjana Martun. Her energy and understanding have always been there to remind me of the human side of the process of research and discovery. Her passion for life and compassion for people are the most admirable bond I have ever seen in a human character.

I especially thank Steve Marsh, Audun Josang, Christian Damsgaard Jensen and Sheikh Mahbub Habib. If there has been a concept of “a professional family”, they would probably have defined it. With their support and professionalism they built a glorious fortress around my scientific morale and enthusiasm, and used their encouragment to shield me from my doubts in key periods of my doctoral research.

I am also immensely grateful to Stuart Clayman. From proofreading my papers to supporting my ups and “Oops”, Stuart’s understanding was always available and highly appreciated.

Finally, I thank my friends and family for their belief and encouragement in “whatever the thing I was doing was”. In particular, I am grateful to my parents for their love, nature and nurture. They instilled in me a sense of responsibility and the mild masochism necessary to pursue my goals. And most of all, to my husband Miroslav and our sons whose support and love were never taken for granted and will always be the core of my emotional cravings.

8 References

Abdul-Rahman, A., & Hailes, S. (1997). A distributed trust model. In *Proceedings of the 1997 workshop on New security paradigms* (pp. 48–60). New York, NY, USA: ACM. http://doi.org/10.1145/283699.283739

Adamic, L., Lauterbach, D., Teng, C.-Y., & Ackerman, M. (2011). {Rating Friends Without Making Enemies}. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 1–8.

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500. http://doi.org/10.2307/1879431

Akoglu, L., Chandy, R., & Faloutsos, C. (2013). Opinion Fraud Detection in Online Reviews by Network Effects. In *Seventh International AAAI Conference on Weblogs and Social Media*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/5981

Alani, H., Kalfoglou, Y., & Shadbolt, N. (2004). Trust strategies for the semantic web. *Proceedings of the Trust, Security and Reputation Workshop at ISWC04*, *7*, 78–85.

Aral, S., & Walker, D. (2012). Identifying Influential and Susceptible Members of Social Networks. *Science*, *337*(6092), 337–341. http://doi.org/10.1126/science.1215842

Ariely, D. (2008). *Predictably Irrational: The Hidden Forces That Shape Our Decisions* (1st ed.). HarperCollins.

Ariely, D., Huber, J., & Wertenbroch, K. (2005). When Do Losses Loom Larger Than Gains? *Journal of Marketing Research*, *42*(2), 134–138. http://doi.org/10.1509/jmkr.42.2.134.62283

Ariely, D., Loewenstein, G., & Prelec, D. (2006). Tom Sawyer and the construction of value. *Journal of Economic Behavior & Organization*, *60*(1), 1–10. http://doi.org/10.1016/j.jebo.2004.10.003

Axelrod, R., & Hamilton, W. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396. http://doi.org/10.1126/science.7466396

Ažderska, T., & Blažič, B. J. (2011). A novel systemic taxonomy of trust in the online environment. In *Proceedings of the 4th European conference on Towards a service-based internet* (pp. 122–133). Berlin, Heidelberg: Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=2050869.2050882

Ažderska, T., & Jerman-Blažič, B. (2012). A Holistic Approach for Designing Human-Centric Trust Systems. *Systemic Practice and Action Research*, 1–34. http://doi.org/10.1007/s11213-012-9259-3

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 65–74). New York, NY, USA: ACM. http://doi.org/10.1145/1935826.1935845

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web* (pp. 519–528). New York, NY, USA: ACM. http://doi.org/10.1145/2187836.2187907

Ba, S. (2001). Establishing online trust through a community responsibility system. *Decision Support Systems*, *31*(3), 323–336. http://doi.org/10.1016/S0167-9236(00)00144-5

Bertalanffy, L. V. (1969). *General System Theory: Foundations, Development, Applications* (Revised). George Braziller, Inc.

Bocan, V., & Fagadar-Cosma, M. (2005). Adaptive Threshold Puzzles. In *The International Conference on Computer as a Tool, 2005. EUROCON 2005* (Vol. 1, pp. 644–647). IEEE. http://doi.org/10.1109/EURCON.2005.1630012

Brown, C. L., & Krishna, A. (2005). *The Skeptical Shopper: A Metacognitive Account for the Effects of Default Options on Choice* (SSRN Scholarly Paper No. ID 2552234). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2552234

Burger, T., Parsons, T., & Toby, J. (1978). Social Systems and the Evolution of Action Theory. *Social Forces*, *57*(1), 308. http://doi.org/10.2307/2577644

Burt, R. S. (2000). *Bandwidth and echo: trust, information, and gossip in social networks*. Chicago, IL: The Author.

Cagnoni, S. (2000). *Real-World Applications of Evolutionary Computing: EvoWorkshops 2000: EvoIASP, EvoSCONDI, EvoTel, EvoSTIM, EvoRob, and EvoFlight, Edinburgh, Scotland, UK, April 17, 2000 Proceedings*. Springer Science & Business Media.

Castelfranchi, C., & Falcone, R. (2000). Trust is much more than subjective probability: Mental components and sources of trust. *32nd Hawaii International Conference on System Sciences - Mini-Track on Software Agents, Maui*, *6*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.105.8007

Castelfranchi, C., & Falcone, R. (2010). *Trust Theory: A Socio-Cognitive and Computational Model*. John Wiley and Sons.

Cattaneo, M. (2011). Belief functions combination without the assumption of independence of the information sources. *International Journal of Approximate Reasoning*, *52*(3), 299–315.

Checkland, P. (1981). *Systems thinking, systems practice*. J. Wiley.

Checkland, P., & Scholes, J. (1999). *Soft systems methodology: a 30-year retrospective*. Wiley.

Chevalier, J. A., & Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, *43*(3), 345–354. http://doi.org/10.1509/jmkr.43.3.345

Chin, S. H. (2009). On application of game theory for understanding trust in networks. In *2009 International Symposium on Collaborative Technologies and Systems* (pp. 106–110). Baltimore, MD, USA. http://doi.org/10.1109/CTS.2009.5067469

Choi, J. A., Koo, M., Choi, I., & Auh, S. (2008). Need for cognitive closure and information search strategy. *Psychology and Marketing*, *25*(11), 1027–1042. http://doi.org/10.1002/mar.20253

Cohen, B. (2003). Incentives Build Robustness in BitTorrent. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.1911

Confer, J. C., Easton, J. A., Fleischman, D. S., Goetz, C. D., Lewis, D. M. G., Perilloux, C., & Buss, D. M. (2010). Evolutionary psychology. Controversies, questions, prospects, and limitations. *The American Psychologist*, *65*(2), 110–126. http://doi.org/10.1037/a0018413

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl, J. (2003). Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 585–592). New York, NY, USA: ACM. http://doi.org/10.1145/642611.642713

David, E., & Jon, K. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. New York, NY, USA: Cambridge University Press.

Davis, J. A. (1967). Clustering and Structural Balance in Graphs. *Human Relations*, *20*(2), 181–187. http://doi.org/10.1177/001872676702000206

Dellarocas, C. (2003). The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms. *Management Science*, *49*(10), 1407–1424. http://doi.org/10.1287/mnsc.49.10.1407.17308

Dellarocas, C., & Wood, C. A. (2008). The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias. *Management Science*, *54*(3), 460 –476. http://doi.org/10.1287/mnsc.1070.0747

Dong, X. L., Berti-Equille, L., & Srivastava, D. (2009). Truth discovery and copying detection in a dynamic world. *Proc. VLDB Endowment*, *2*(1), 562–573.

Doreian, P., & Krackhardt, D. (2001). Pre-transitive balance mechanisms for signed networks*. *The Journal of Mathematical Sociology*, *25*(1), 43–67. http://doi.org/10.1080/0022250X.2001.9990244

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *The Academy of Management Review*, *14*(4), 532. http://doi.org/10.2307/258557

Fehr, E., & Gächter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=229149

Fisher, I. (2011). *The Money Illusion*. Kessinger Publishing, LLC.

Flynn, F. J., Reagans, R. E., & Guillory, L. (2010). Do you two know each other? Transitivity, homophily, and the need for (network) closure. *Journal of Personality and Social Psychology*, *99*(5), 855–869. http://doi.org/10.1037/a0020961

Flyvbjerg, B. (2006). Five Misunderstandings About Case-Study Research. *Qualitative Inquiry*, *12*(2), 219–245. http://doi.org/10.1177/1077800405284363

Fortuna, B., Grobelnik, M., & Mladenic, D. (2007). OntoGen: Semi-automatic Ontology Editor. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.6909

Fox, C. R., & Clemen, R. T. (2005). Subjective Probability Assessment in Decision Analysis: Partition Dependence and Bias Toward the Ignorance Prior. *Management Science*, *51*(9), 1417–1432. http://doi.org/10.1287/mnsc.1050.0409

Fudenberg, D., & Tirole, J. (1991). *Game Theory*. The MIT Press.

Fukuyama, F. (1996). *Trust: The Social Virtues and The Creation of Prosperity* (1st Free Press Pbk. Ed edition). New York: Free Press.

Gallos, L. K., Rybski, D., Liljeros, F., Havlin, S., & Makse, H. A. (2012). How People Interact in Evolving Online Affiliation Networks. *Physical Review X*, *2*(3), 031014. http://doi.org/10.1103/PhysRevX.2.031014

Gambetta, D. (1988). Can We Trust Trust? *Trust: Making and Breaking Cooperative Relations*, 213–237.

Gary S. Becker, & Kevin M. Murphy. (2003). *Social Economics: Market Behavior in a Social Environment*. Retrieved from http://www.hup.harvard.edu/catalog.php?isbn=9780674011212

Goel, S., Watts, D. J., & Goldstein, D. G. (2012). The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce* (pp. 623–638). New York, NY, USA: ACM. http://doi.org/10.1145/2229012.2229058

Golbeck, J. (2006). Using Trust and Provenance for Content Filtering on the Semantic Web. *Proceedings of the Workshop on Models of Trust on the Web, at the 15th WWW Conference*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.108.1121

Granovetter, M. (1978). Threshold Models of Collective Behavior. *American Journal of Sociology*, *83*(6), 1420–1443. http://doi.org/10.1086/226707

Granovetter, M. (1983). The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, *1*(1983), 201–233. http://doi.org/10.2307/202051

Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a Citation Network. *BMJ*, *339*, b2680–b2680. http://doi.org/10.1136/bmj.b2680

Grossi, D., Dignum, F., & Meyer, J.-J. C. (2005). Contextual Taxonomies. In J. Leite & P. Torroni (Eds.), *Computational Logic in Multi-Agent Systems* (pp. 33–51). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/11533092_3

Heath, T., Motta, E., & Petre, M. (2007). Computing Word-of-Mouth Trust Relationships in Social Networks from Semantic Web and Web2.0 Data Sources. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.96.5231

Heider, F. (1982). *The Psychology of Interpersonal Relations*. Routledge.

Holland, P. W., & Leinhardt, S. (1974). *The Statistical Analysis of Local Structure in Social Networks* (Working Paper No. 44). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w0044

Hsee, C. K. (1998). Less is better: when low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making*, *11*(2), 107–121. http://doi.org/10.1002/(SICI)1099-0771(199806)11:2<107::AID-BDM292>3.0.CO;2-Y

Huh, Y. E., Vosgerau, J., & Morewedge, C. K. (2014). Social Defaults: Observed Choices Become Choice Defaults. *Journal of Consumer Research*, *41*(3), 746–760. http://doi.org/10.1086/677315

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped Distribution of Product Reviews. *Communications of ACM*, *52*(10), 144–147. http://doi.org/10.1145/1562764.1562800

Huynh, T. D. (2006, June). Trust and Reputation in Open Multi-Agent Systems [Thesis]. Retrieved July 21, 2011, from http://eprints.ecs.soton.ac.uk/12759/

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*(6), 995–1006. http://doi.org/10.1037/0022-3514.79.6.995

Johnson, E. J., Shu, S. B., Dellaert, B. G. C., Fox, C., Goldstein, D. G., Häubl, G., … Weber, E. U. (2012). Beyond nudges: Tools of a choice architecture. *Marketing Letters*, *23*(2), 487–504. http://doi.org/10.1007/s11002-012-9186-1

Jordan, N. (2003). *Themes in Speculative Psychology*. Routledge.

Jøsang, A. (2001). A logic for uncertain probabilities. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, *9*(3), 279–311. http://doi.org/10.1016/S0218-4885(01)00083-1

Josang, A. (2008). Conditional Reasoning with Subjective Logic [Journal Article]. Retrieved November 27, 2012, from http://eprints.qut.edu.au/14842/

Josang, A. (2013). *Subjective Logic* (Draft). Retrieved from http://folk.uio.no/josang/papers/subjective_logic.pdf

Jøsang, A., Ažderska, T., & Marsh, S. (2012). Trust Transitivity and Conditional Belief Reasoning. In T. Dimitrakos, R. Moona, D. Patel, & D. McKnight (Eds.), *Trust Management VI* (Vol. 374, pp. 68–83). Springer Boston. Retrieved from http://www.springerlink.com/content/y7467n0n1177753u/abstract/

Josang, A., Daniel, M., & Vannoorenberghe, P. (2003). Strategies for combining conflicting dogmatic beliefs. In *Proceedings of the Sixth International Conference of Information Fusion, 2003* (Vol. 2, pp. 1133 – 1140). http://doi.org/10.1109/ICIF.2003.177365

Jøsang, A., & Elouedi, Z. (2007). Interpreting Belief Functions as Dirichlet Distributions. In K. Mellouli (Ed.), *Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (pp. 393–404). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-75256-1_36

Josang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision Support Systems*, *43*(2), 618–644. http://doi.org/10.1016/j.dss.2005.05.019

Jøsang, A., Marsh, S., & Pope, S. (2006). Exploring different types of trust propagation. In *Proceedings of the 4th international conference on Trust Management* (pp. 179–192). Berlin, Heidelberg: Springer-Verlag. http://doi.org/10.1007/11755593_14

Josang, A., & O'Hara, S. (2010). The base rate fallacy in belief reasoning. In *2010 13th Conference on Information Fusion (FUSION)* (pp. 1–8). IEEE.

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review*, *93*(5), 1449–1475. http://doi.org/10.1257/000282803322655392

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. http://doi.org/10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–91.

Kamvar, S. D., Schlosser, M. T., & Garcia-Molina, H. (2003). The Eigentrust algorithm for reputation management in P2P networks. In *Proceedings of the twelfth international conference on World Wide Web - WWW '03* (p. 640). Budapest, Hungary. http://doi.org/10.1145/775152.775242

Kleinberg, J. (2000). The Small-World Phenomenon: An Algorithmic Perspective. *In Proceedings of the 32nd ACM Symposium on Theory of Computing*, 163–170.

Klos, T., & La Poutré, H. (2006). A versatile approach to combining trust values for making binary decisions. In *Proceedings of the 4th international conference on Trust Management* (pp. 206–220). Berlin, Heidelberg: Springer-Verlag. http://doi.org/10.1007/11755593_16

Krackhardt, D. (1992). The Strength of Strong Ties: The Importance of Philos in Organizations. In N. Nohria & R. Eccles (Eds.), *Networks and Organizations: Structure, Form, and Action* (pp. 216–239). Harvard Business School Press.

Kramer, M. (2007). Self-Selection Bias in Reputation Systems. In *Trust Management* (Vol. 238, pp. 255–268). Springer Boston. Retrieved from http://dx.doi.org/10.1007/978-0-387-73655-6_17

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing." *Psychological Review*, *103*(2), 263–283. http://doi.org/10.1037/0033-295X.103.2.263

Kwei-Jay Lin, Haiyin Lu, Tao Yu, & Chia-en Tai. (2005). A reputation and trust management broker framework for Web applications. In *The 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service, 2005. EEE '05. Proceedings* (pp. 262– 269). IEEE. http://doi.org/10.1109/EEE.2005.14

Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010). Signed networks in social media. In *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 1361–1370). New York, NY, USA: ACM. http://doi.org/10.1145/1753326.1753532

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., & Hurst, M. (2007). Cascading Behavior in Large Blog Graphs. *Proceedings of 7th SIAM International Conference on Data Mining (SDM)*. Retrieved from http://arxiv.org/abs/0704.2803

Lieberman, M. B., & Montgomery, D. B. (1988). First-mover advantages. *Strategic Management Journal*, *9*(S1), 41–58. http://doi.org/10.1002/smj.4250090706

Lou, T., Tang, J., Hopcroft, J., Fang, Z., & Ding, X. (2013). Learning to Predict Reciprocity and Triadic Closure in Social Networks. *ACM Transactions on Knowledge Discovery Data*, *7*(2), 5:1–5:25. http://doi.org/10.1145/2499907.2499908

Luhmann, N. (1979). *Trust and power: two works*. Wiley.

Luhmann, N. (1987). *Soziale Systeme: Grundriß einer allgemeinen Theorie* (Auflage: 1.). Frankfurt am Main: Suhrkamp Verlag.

Luhmann, N. (2005). *Risk: a sociological theory*. Transaction Publishers.

Lynch, J. G., & Ariely, D. (2000). Wine Online: Search Costs Affect Competition on Price, Quality, and Distribution. *Marketing Science*, *19*(1), 83–103. http://doi.org/10.1287/mksc.19.1.83.15183

MacMullen, R. (2003). *Feelings in history, ancient and modern*. Regina Books.

Makridakis, S., & Taleb, N. (2009). Decision making and planning under low levels of predictability. *International Journal of Forecasting*, *25*(4), 716–733. http://doi.org/16/j.ijforecast.2009.05.013

Marsh, S. (1994). Formalising trust as a computational concept, 170.

Marsh, S., & Briggs, P. (2009). Examining Trust, Forgiveness and Regret as Computational Concepts. In J. Golbeck (Ed.), *Computing with Social Trust* (pp. 9–43). London: Springer London. Retrieved from http://www.springerlink.com/index/10.1007/978-1-84800-356-9_2

Marti, S., & Garciamolina, H. (2006). Taxonomy of trust: Categorizing P2P reputation systems☆. *Computer Networks*, *50*(4), 472–484. http://doi.org/10.1016/j.comnet.2005.07.011

Marti, S., Giuli, T. J., Lai, K., & Baker, M. (2000). Mitigating routing misbehavior in mobile ad hoc networks. In *Proceedings of the 6th annual international conference on Mobile computing and networking* (pp. 255–265). New York, NY, USA: ACM. http://doi.org/10.1145/345910.345955

McDermott, R., Fowler, J. H., & Smirnov, O. (2008). On the Evolutionary Origin of Prospect Theory Preferences. *The Journal of Politics*, *70*(02), 335–350. http://doi.org/10.1017/S0022381608080341

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444. http://doi.org/10.1146/annurev.soc.27.1.415

Meulpolder, M., Pouwelse, J. A., Epema, D. H. J., & Sips, H. J. (2009a). BarterCast: A Practical Approach to Prevent Lazy Freeriding in P2P Networks. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.153.578

Meulpolder, M., Pouwelse, J. A., Epema, D. H. J., & Sips, H. J. (2009b). BarterCast: A practical approach to prevent lazy freeriding in P2P networks. In *IEEE International Symposium on Parallel Distributed Processing, 2009. IPDPS 2009* (pp. 1 –8). http://doi.org/10.1109/IPDPS.2009.5160954

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., … Alon, U. (2004). Superfamilies of Evolved and Designed Networks. *Science*, *303*(5663), 1538–1542. http://doi.org/10.1126/science.1089167

Mishra, A., & Bhattacharya, A. (2011). Finding the bias and prestige of nodes in networks based on trust scores. In *Proceedings of the 20th international conference on World wide web* (pp. 567–576). New York, NY, USA: ACM. http://doi.org/10.1145/1963405.1963485

Mui, L., Mohtashemi, M., Ang, C., Szolovits, P., & Halberstadt, A. (2001). Ratings in Distributed Systems: A Bayesian Approach. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.9051

Myers, S. A., Zhu, C., & Leskovec, J. (2012). Information Diffusion and External Influence in Networks. *arXiv:1206.1331*. Retrieved from http://arxiv.org/abs/1206.1331

Newman, M. E. J. (2000). Models of the Small World - A Review. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.3390

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM REVIEW*, *45*, 167–256.

Noorian, Z., & Ulieru, M. (2010). The State of the Art in Trust and Reputation Systems: A Framework for Comparison. *Journal of Theoretical and Applied Electronic Commerce Research*, *5*(2). http://doi.org/10.4067/S0718-18762010000200007

Odum, E., & Barrett, G. W. (2004). *Fundamentals of Ecology* (5th ed.). Brooks Cole.

Pennington, R., Wilcox, H. D., & Grover, V. (2003). The Role of System Trust in Business-to-Consumer Transactions. *J. Manage. Inf. Syst.*, *20*(3), 197–226.

Pfister, H., & Bohm, G. (2008). The multiplicity of emotions: a framework of emotional functions in decision-making. *Judgment and Decision Making*, *3*(1), 5–17.

Pohl, R. F. (2012). *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. Psychology Press.

Pope, S. (2005). Analysis of Competing Hypotheses using Subjective Logic. *Systems Technology*, (June), 13–16.

Pouwelse, J. A., Garbacki, P., Wang, J., Bakker, A., Yang, J., Iosup, A., … Sips, H. J. (2008). TRIBLER: a social-based peer-to-peer system. *Concurrency and Computation: Practice and Experience*, *20*(2), 127–138. http://doi.org/10.1002/cpe.1189

Powers, W. T. (2005). *Behavior: The Control Of Perception* (2 Rev Exp). Benchmark Publications, Inc.

Resnick, P., & Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of eBay' s reputation system. In *Advances in Applied Microeconomics* (Vol. 11, pp. 127–157). Bingley: Emerald (MCB UP). Retrieved from http://www.emeraldinsight.com/books.htm?chapterid=1782046

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to Recommender Systems Handbook. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 1–35). Boston, MA: Springer US. Retrieved from http://www.pubzone.org/dblp/reference/rsh/RicciRS11

Ries, S. (2007). Certain trust: a trust model for users and agents. In *Proceedings of the 2007 ACM symposium on Applied computing* (pp. 1599–1604). New York, NY, USA: ACM. http://doi.org/10.1145/1244002.1244342

Sandberg, O. (2005). *Searching in a Small World*. Chalmers University of Technology and G¨oteborg University, G¨oteborg, Sweden. Retrieved from ISSN 1652-9715.

Savani, M. (2014). The Behavioral Foundations of Public Policy. *Political Studies Review*, *12*(2), 308–309. http://doi.org/10.1111/1478-9302.12053_97

Schneier, B. (2012). *Liars and outliers: enabling the trust that society needs to thrive*. Indianapolis, In: Wiley.

Schwöbbermeyer, H. (2008). Network Motifs. In B. H. Junker & F. Schreiber (Eds.), *Analysis of Biological Networks* (pp. 85–111). John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9780470253489.ch5/summary

Shafer, G. (1976). *Mathematical Theory of Evidence*. Princeton University Press

Shu, S. B. (2008). Future-biased search: the quest for the ideal. *Wiley*, *21*(4), 352–377. http://doi.org/ZDB-ID 2277190

Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: do people adjust from provided anchors?

Journal of Personality and Social Psychology, *99*(6), 917–932. http://doi.org/10.1037/a0021540

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. http://doi.org/10.2307/1884852

Smets, P. (2007). Analyzing the combination of conflicting belief functions. *Inf. Fusion*, *8*(4), 387–412. http://doi.org/10.1016/j.inffus.2006.04.003

Smith, N. C., Goldstein, D. G., & Johnson, E. J. (2013). Choice Without Awareness: Ethical and Policy Implications of Defaults. *Journal of Public Policy & Marketing*, *32*(2), 159–172. http://doi.org/10.1509/jppm.10.114

Sunstein, C. R., & Thaler, R. H. (2009). *Nudge: Improving Decisions About Health, Wealth and Happiness*. London: Penguin.

Surowiecki, J. (2005). *The Wisdom Of Crowds*. Anchor Books.

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian Paternalism. *American Economic Review (AER)*, *93*(3), 175–179. http://doi.org/10.1257/000282803321947001

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. http://doi.org/10.1126/science.185.4157.1124

Victor, P., Cornelis, C., De Cock, M., & Herrera-Viedma, E. (2011). Practical aggregation operators for gradual trust and distrust. *Fuzzy Sets and Systems*, *184*(1), 126–147. http://doi.org/10.1016/j.fss.2010.10.015

Wang, X. T., Simons, F., & Brédart, S. (2001). Social cues and verbal framing in risky choice. *Journal of Behavioral Decision Making*, *14*(1), 1–15. http://doi.org/10.1002/1099-0771(200101)14:1<1::AID-BDM361>3.0.CO;2-N

Wang, Y., & Singh, M. P. (2010). Evidence-based trust: A mathematical model geared for multiagent systems. *ACM Transactions on Autonomous and Adaptive Systems*, *5*(4), 14:1–14:28. http://doi.org/10.1145/1867713.1867715

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Watts, D. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, *99*(9), 5766–5771. http://doi.org/10.1073/pnas.082090499

Wernicke, S., & Rasche, F. (2006). FANMOD: A Tool for Fast Network Motif Detection. *Bioinformatics*, *22*(9), 1152–1153. http://doi.org/10.1093/bioinformatics/btl038

Wikipedia. (2014, August 29). Social network analysis. In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/w/index.php?title=Social_network_analysis&oldid=621187763

Wouter de Nooy, Mrvar, A., & Batagelj, V. (2012). *Exploratory Social Network Analysis with Pajek* (2nd ed.). Cambridge University Press. Retrieved from http://www.cambridge.org/si/knowledge/isbn/item6436521/Exploratory%20Social%20Network%20Analysis%20with%20Pajek/?site_locale=sl_SI

Xiong, L., & Liu, L. (2002). Building Trust in Decentralized Peer-to-Peer Electronic Communities. *In the 5th International Conference on Electronic Commerce Research.* Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.6197

Yardi, S., & Boyd, D. (2010). Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter. *Bulletin of Science, Technology & Society*, *30*(5), 316 –327. http://doi.org/10.1177/0270467610380011

Yin, R. K. (1981). The Case Study Crisis: Some Answers. *Administrative Science Quarterly*, *26*(1), 58. http://doi.org/10.2307/2392599

Yuan, W., Hong, Y., & Pavlou, P. (2012). Do Consumers Trust Online Product Reviews? An Experimental Study of Biases in Online Product Reviews. *AMCIS 2012*

Proceedings. Retrieved from http://aisel.aisnet.org/amcis2012/proceedings/HCIStudies/27

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151–175. http://doi.org/37/0003-066X.35.2.151

Zaltzman, O., & Le, Q. D. (2007). Trust Based Moderation. Retrieved from http://www.google.com/patents/about/11_941_009_TRUST_BASED_MODERATION.html?id=SgjHAAAAEBAJ

Zhang, J. (2011). Anchoring effects of recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 375–378). New York, NY, USA: ACM. http://doi.org/10.1145/2043932.2044010

Zhao, B., Rubinstein, B. I. P., Gemmell, J., & Han, J. (2012). A Bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endowment*, *5*(6), 550–561.

Zhou, R., & Hwang, K. (2007). PowerTrust: A Robust and Scalable Reputation System for Trusted Peer-to-Peer Computing. *IEEE Transactions on Parallel and Distributed Systems*, *18*(4), 460–473. http://doi.org/10.1109/TPDS.2007.1021

Index of Figures

Figure 1: *a) Example of balanced triads b) Example of unbalanced triads.* 10
Figure 2: *Transitive trust fusion.* 18
Figure 3: *The interdependencies between active entity behaviour and the trust system's subcomponents.* 31
Figure 4: *The interdependencies between context and the trust system's subcomponents.* 33
Figure 5: *The interdependencies between resilience and evolutionism and the trust system's subcomponents* 34
Figure 6: *The interdependencies between dynamics and the trust system's subcomponents.* 35
Figure 7: *Ontological map of the ACM DL query "trust reputation systems".* 36
Figure 8: *Evaluation of the interdependencies between active entity behaviour and BarterCast's subcomponents.* 40
Figure 9: *Evaluation of the interdependencies between context and BarterCast's subcomponents.* 41
Figure 10: *Evaluation of the interdependencies between resilience and evolutionism and BarterCast's subcomponents.* 42
Figure 11: *Evaluation of the interdependencies between time and system dynamics and BarterCast's subcomponents.* 42
Figure 12: *Evaluation of the interdependencies between active entity behaviour and Yahoo! Answer's subcomponents.* 44
Figure 13: *Evaluation of the interdependencies between context and Yahoo! Answer's subcomponents.* 45
Figure 14: *Evaluation of the interdependencies between resilience and evolutionism and Yahoo! Answer's subcomponents.* 46
Figure 15: *Evaluation of the interdependencies between time and system dynamics and Yahoo! Answer's subcomponents.* 47
Figure 16: *Rating distribution for each review by offline users a) numerically b) descriptively.* 59
Figure 17: *Rating distribution for the two reviews for a) survey 1, b) survey 2 and c) survey 3.* 59
Figure 18: *Difference in rating distribution between surveys 1 and 2 for a) review 1 b) review 2.* 60
Figure 19: *Difference in rating distribution between surveys 1 and 3 for a) review 1 b) review 2.* 61
Figure 20: *Difference in rating distribution between surveys 2 and 3 for a) review 1 b) review 2.* 61

Figure 21: *Difference in ratings between the same options in the two descriptive surveys.* 62
Figure 22: *Representation of all the possible triads that can be formed model.* 72
Figure 23: *The data sets triad significance profiles.* 74
Figure 24: *Balance theory predictive performance on each of the data sets:* 1 = Epinions, 2 = Slashdot, 3 = Wikipedia. According to BT's predictions, the colour that should be dominant on the graphs is: a) blue; b) red; c) red; d) blue 76
Figure 25: *Separate triad-type analysis of the closures in the Wikipedia data set.* Included in the analysis are also the neutral opinions, and excluded is the (+,+,+) triad for statistical reasons. 77
Figure 26: *The (dynamics of the) process of a triad formation.* 79
Figure 27: *a) Opinion dependence b) opinion conflict.* 81
Figure 28: a) *Subjectivization of 1/0/-1 sentiments to opinions; b) desubjectivization of opinions to 1/0/-1 sentiments.* 82
Figure 29: *Comparative analysis of the performance of balance theory, subjective logic and the bias reduction method through the ROC curves.* 89
Figure 30*: a) Opinion dependence b) opinion conflict.* 90

Index of Tables

Table 1: *Organizing principles of Jordan's taxonomy.* The categories to which trust systems are assigned are italicized. 27
Table 2: *Breakdown of the trust system components (Marti and Garciamolina, 2006).* 30
Table 3: *Statistics about the experimental setting.* 58
Table 4: *Mean value for the ratings of the surveys.* 59
Table 5: *Closed triads classification.* 72
Table 6: *Number of triads per type.* 72
Table 7: *Performance of balance theory as a predictive model* 76
Table 8: *Separate sign closure for each triad type in each of the data sets*. The bolded red numbers denote BT's sign prediction for that triadic closure. 76
Table 9: *Separate triad-type analyses of the triadic closures in Wikipedia's data set* (included are also the ones that contain neutral opinion). Bolded are the results where BT is expected to hold. 77
Table 10: *SL operators' performance on the Slashdot data set.* 83
Table 11: *SL operators' performance on the Epinions data set.* 83
Table 12: *SL operators' performance on the Wikipedia data set (without accounting for the zeros).* 83
Table 13: *Wikipedia aggregated opinion-to-value representations and the SL operators' performance if neutral opinions are taken into account.* The subjective values for the sentiments are: - = [0.1.0.8 0.1 0.5], + = [0.8 0.1 0.1 0.5], 0 = [0.45 0.45 0.1 0.5]. 84
Table 14: *SL operators' performance on the Slashdot data set after applying the bias reduction method.* 87
Table 15: *SL operators' performance on the Wikipedia data set after applying the bias reduction method.* 87

Index of Algorithms

9.2.1 Algorithm 1: The Pseudo-code for the Extraction of Transitive Triads from the Analysed Data Sets..... 119

9 Appendix

9.1 Chapter 4: User Bias Identification in Online Trust Systems

Survey Questions (Reviews)

This appendix shows the reviews that were used in the surveys. They can also be found at http://www.epinions.com/reviews/hewlett-packard-probook-4530s-15-6-i3-2310m-xu015utaba-pc-notebook

Review 1:

User Rating: OK; **Ease of Use:** 2; **Quality of Tech Support:** 1

Pros:1) Intel core i3 processor; 2) Finger Print Sensor; 3) Battery life

Cons:1) Build material; 2) Intel integrated graphics; 3) 320 GB hard disk space

The Bottom Line: This laptop is not recommended for anyone, because of poor support quality.

HP products were good three years ago. People loved them for their looks, reliability, performance and budget price. But since 2008, the reliability ratio of HP compared to its competitors has been far lower in almost all aspects including price, performance, looks and, most importantly, reliability. I myself was a big fan of this company, but after facing the failure of two consecutive PCs from this company, I changed my view. My wife had this laptop and I bought it for her as her birthday present three months ago, but after just two months, the screen got dead spot (small black spots) and the hinge connecting the screen with the keyboard also got broken. Still, the laptop was running fine, but last week its keyboard also stopped working, which sucks. All the money I paid for it gone in vain. My experience about: PROCESSOR: Intel core i3 processor clocked at 2.1 GHZ works really very well for multitasking but it is not designed to handle more tasks. This PC is fine for web surfing, word processing, office work, watching movies. GRAPHICS: Intel integrated graphics are not capable of running blue ray movies silky smooth. BUILD MATERIAL: This laptop is made of seriously very cheap plastic. Its glossy surface is just a fingerprint magnet and you have to clean up the laptop after every single use. The glossy screen also causes panic while watching movies in sunlight. LACK OF HDMI PORT: This laptop lacks one of the most important ports included in the laptops of this price range, i.e. an HDMI port. The sound quality is good, battery life is also impressive and lasts about three hours (6 cells) even when watching movies. The security features like Finger Print Sensor work great. Its light weight makes it easier to carry this laptop everywhere but it can't be called an ultraportable laptop. Overall, this laptop isn't recommended to anyone, whether the person is a student or housewife.

[**Recommended**: No]

Review 2:

User Rating: Excellent; **Ease of Use:** 5; **Quality of Tech Support:** 4

Pros: I would buy another computer from HP

Cons: Multiple hardware faults (e.g. hard to type and sensitive mouse pad)

The HP ProBook 4530s is an extremely good product. The computer is lightweight and rugged. It's able to be put in a backpack and stuffed wherever you need to go. When I bought and received the computer I liked how the computer didn't come with a lot of extra software junk. Next, I like having a numeric pad and I will continue to buy computers with them from now on. For just web browsing and playing light computer games this computer is extremely fast. The keypad is not sensitive enough and I have to hit the keys hard a lot. Also if you have multiple fingers close to the mouse pad, the mouse pad goes crazy due to its many settings. The speaker system is excellent, the loudest system I have heard on any computer system. The software for the finger ID gets weird every now and then and I have to input my fingers a couple of times to log in. The settings are easy to change on the computer. The battery life is average compared to other computers. The computer hardware is nicely laid out and easy to find. The computer is worth the price I paid for it.

[**Recommended**: Yes]

9.2 Chapter 5: A Formal Model for Alleviating User Bias in Computational Trust Systems

9.2.1 Algorithm 1: The Pseudo-code for the Extraction of Transitive Triads from the Analysed Data Sets

```
fun produceTriplets(neighbours, resultsFile, signsFile)
{
   for nodeNum=0..neighbours.size()-1 {
      adjacent = neighbours[nodeNum];

      for j=0..adjacent.size()-1 {
         for b=0..adjacent.size()-1 {
            // skip processing for itself
            if (adjacent[j].nodeId==adjacent[b].nodeId) continue

            otherAdjacent = neighbours[adjacent[b].nodeId-1]
            // Check if j is also adjacent to the b neighbours
            // find returns -1 if not found otherwise an
            // index in the first parameter vector
            pos = find(otherAdjacent, adjacent[j].nodeId)
            if (pos != -1) { // if found
               resultsFile.writeLine(nodeNum+1, adjacent[b].nodeId,
                  adjacent[j].nodeId, adjacent[b].sign,
                  otherAdjacent[pos].sign, adjacent[j].sign)

               signsFile.writeLine(adjacent[b].sign,
```

```
                otherAdjacent[pos].sign, adjacent[j].sign) }
    } } } }
```

---

neighbours is a vector containing vectors of <*nodeId*, *sign*> pairs

For a given node *n*, its adjacent nodes and signs pair can be found at *neighbours[n-1]*

resultsFile is the output file where each line contains a tuple of size 6 for each triplet. The first three elements represent the node ids, the last three elements represent the signs (votes) for those node ids

signsFile is the output file in which each line contains the latter triplet from *resultsFile* (see previous comment)

9.2.2 Balance Theory (neutral opinion analysis of the Wikipedia data set)

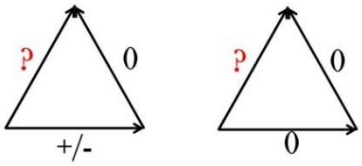


Figure A.1: Aim of our analysis: investigating triadic closure in the presence of neutral opinion.

From the results obtained, we come to the following conclusions:

1. The general tendency of relationship formation is towards expressing positive attitudes, which only supports the previously detected positivity bias.
2. When a positive edge is present in the triad (together with a neutral), it is more likely that the new closure will be positive (Figure 2).



Figure A.2: When a positive edge is present in the triad (together with a neutral), it is more likely that the new closure will be positive.

However, it is more likely to befriend someone towards whom your friend is ambivalent than someone with whom your ambivalent connection is a friend.

3. When there is an option to express neutral opinion (i.e. ambivalence), the statistical significance of unbalanced triads greatly increases.
4. When a negative edge is present in the triad (enemy relationship), we see a more or less equal likelihood of both positive and negative closure (Figure 3).

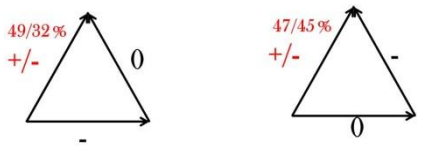


Figure A.3: When a negative edge is present in the triad (enemy relationship), we see a more or less equal likelihood of both positive and negative closure.

However, it is more likely to be friends than to be enemies with someone towards whom your enemy is ambivalent (49>32 and 47>45), **and** it is slightly more likely to be friends with the ambivalent partner of your enemy than with the enemy of your ambivalent partner (49>47).

1. (+, 0) is closer to (+, +) than (+, -): a neutral foaf relationship is more likely to replicate my friend's positive attitudes than his negative attitudes
2. (0, +) is closer to (-, +) than (+, +): a neutral connection is more likely to resemble my tendency of assigning enmity rather than friendship
3. (-, 0) is closer to (-, +) than (-, -): a neutral friend of my enemy is more likely to replicate my enemy's positive attitudes than his negative ones
4. (0, -) is closer to (-, -) than (+, -): a neutral connection is more likely to resemble my tendency to assign enmity rather than friendship (same as in result 2)

9.2.3 Subjective Logic: Sensitivity tests (if no neutral opinions are taken into account) – Slashdot Data Set

I) w_pos = [0.8 0.1 0.1 0.5]; w_neg = [0.1 0.8 0.1 0.5]

	Opposite Belief Favouring	Uncertainty Favouring	Base-rate Sensitivity
(- o +)	[0.16 0.65 0.19 0.5] = -	[0.08 0.01 0.91 0.5] = 0	[0.12 0.015 0.865 0.5] = 0
(- o -)	[0.65 0.16 0.19 0.5] = +	[0.01 0.08 0.91 0.5] = 0	[0.015 0.12 0.865 0.5] = 0
(+ o -)	[0.16 0.65 0.19 0.5] = -	[0.08 0.64 0.28 0.5] = -	[0.085 0.68 0.235 0.5] = -
(+ o +)	[0.65 0.16 0.19 0.5] = +	[0.64 0.08 0.28 0.5] = +	[0.68 0.085 0.235 0.5] = +

II) w_pos = [0.7 0.2 0.1 0.5]; w_neg = [0.2 0.7 0.1 0.5];

	Opposite Belief Favouring	Uncertainty Favouring	Base-rate Sensitivity
(- o +)	[0.28 0.53 0.19 0.5] = -	[0.14 0.04 0.82 0.5] = 0	[0.175 0.05 0.775 0.5] = 0
(- o -)	[0.53 0.28 0.19 0.5] = +	[0.04 0.14 0.82 0.5] = 0	[0.05 0.175 0.775 0.5] = 0
(+ o -)	[0.28 0.53 0.19 0.5] = -	[0.14 0.49 0.37 0.5] = -	[0.15 0.525 0.325 0.5] = -
(+ o +)	[0.53 0.28 0.19 0.5] = +	[0.49 0.14 0.37 0.5] = +	[0.525 0.15 0.325 0.5] = +

III) w_pos = [0.6 0.3 0.1 0.5]; w_neg = [0.3 0.6 0.1 0.5];

	Opposite Belief Favouring	Uncertainty Favouring	Base-rate Sensitivity
(- o +)	[0.36 0.45 0.19 0.5] = 0	[0.18 0.09 0.73 0.5] = 0	[0.21 0.105 0.685 0.5] = 0
(- o -)	[0.45 0.36 0.19 0.5] = 0	[0.09 0.18 0.73 0.5] = 0	[0.105 0.21 0.685 0.5] = 0
(+ o -)	[0.36 0.45 0.19 0.5] = 0	[0.18 0.36 0.46 0.5] = -	[0.195 0.39 0.415 0.5] = -
(+ o +)	[0.45 0.36 0.19 0.5] = 0	[0.36 0.18 0.46 0.5] = +	[0.39 0.195 0.415 0.5] = +

IV) w_pos = [0.5 0.4 0.1 0.5]; w_neg = [0.4 0.5 0.1 0.5];

	Opposite Belief Favouring	Uncertainty Favouring	Base-rate Sensitivity

(- o +)	[0.4 0.41 0.19 0.5] = 0	[0.2 0.16 0.64 0.5] = /	[0.225 0.18 0.595 0.5] = /
(- o -)	[0.41 0.4 0.19 0.5] = 0	[0.16 0.2 0.64 0.5] = /	[0.18 0.2250.595 0.5] = /
(+ o -)	[0.4 0.41 0.19 0.5] = 0	[0.2 0.25 0.55 0.5] = /	[0.22 0.275 0.505 0.5] = /
(+ o +)	[0.41 0.4 0.19 0.5] = 0	[0.25 0.2 0.55 0.5] = /	[0.275 0.22 0.505 0.5] = /

V) w_pos = [0.45 0.45 0.1 0.5]; w_neg = [0.45 0.45 0.1 0.5];

	Opposite Belief Favouring	Uncertainty Favouring	Base-rate Sensitivity
(- o +)	[0.405 0.405 0.19 0.5] = 0	[0.2025 0.2025 0.55 0.5] = /	[0.225 0.225 0.55 0.5] = /
(- o -)	[0.405 0.405 0.19 0.5] = 0	[0.2025 0.2025 0.55 0.5] = /	[0.225 0.225 0.55 0.5] = /
(+ o -)	[0.405 0.405 0.19 0.5] = 0	[0.2025 0.2025 0.55 0.5] = /	[0.225 0.225 0.55 0.5] = /
(+ o +)	[0.405 0.405 0.19 0.5] = 0	[0.2025 0.2025 0.55 0.5] = /	[0.225 0.225 0.55 0.5] = /

1. Unlike Balance theory, SL operators are able to express and work with ambivalent opinions = through the *uncertainty* parameter in the SL opinions
2. The extent of belief mass can be utilized to decide on the nature of uncertainty: whether it leans towards belief or towards disbelief
3. Both the uncertainty favouring and the base-rate sensitivity operators produce overly doubtful opinions, which is inhibiting for the interactions
4. We find the nodes with high belief by the community members; the results show that this number is slightly higher (398,904) than the number of overall positives (392,326). However, this does not mean that the same nodes that are considered positive in the binary domain are treated with belief in the subjective domain.
5. All of the operators are extremely sensitive to changes in the belief mass. The amount of uncertainty increases greatly with any change, leading to difficulties in the decision-making process. **The bias reduction method deals with this issue**.

Bibliography

Publications Related to the Thesis

Journal Articles

Ažderska, T., & Jerman-Blažič, B. (2012). A Holistic Approach for Designing Human-Centric Trust Systems. *Systemic Practice and Action Research*, 1–34. http://doi.org/10.1007/s11213-012-9259-3

Ažderska, T., & Blažic, B. J. (2011). A novel systemic taxonomy of trust in the online environment. In *Proceedings of the 4th European conference on Towards a service-based internet* (pp. 122–133). Berlin, Heidelberg: Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=2050869.2050882

Ažderska, T., & Blažič, B. J. (2012). Trust as an Organismic Trait of E-Commerce Systems. In G. Quirchmayr, J. Basl, I. You, L. Xu, & E. Weippl (Eds.), *Multidisciplinary Research and Practice for Information Systems* (pp. 161–175). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-32498-7_13

Conference Papers

Ažderska, T., & Jerman Blažič, B. (2010). Providing trust and reputation in P2P network. In *Zbornik devetnajste mednarodne Elektrotehniške in računalniške konference ERK 2010* (Vol. zv. B, pp. 139–142). Portorož, Slovenija: IEEE, Zajc, Baldomir.

Ažderska, T., & Jerman Blažič, B. (2011b). Online trust and reputation : towards socially aware taxonomy of trust (pp. 79–85). Presented at the 1st UNITE doctoral symposium, Bucharest: Jardim-Gonçalves, Ricardo, Doumeingts, Guy, Stanescu, Aurelian Mihai.

Ažderska, T., & Jerman Blažič, B. (2012). Developing Trust and Reputation Taxonomy for a Dynamic Network Environment (pp. 109–114). ICN 2012, *The Eleventh International Conference on Networks*. Retrieved from http://www.thinkmind.org/index.php?view=article&articleid=icn_2012_5_40_10173

Ažderska, T. (2012). Co-evolving Trust Mechanisms for Catering User Behavior. In T. Dimitrakos, R. Moona, D. Patel, & D. McKnight (Eds.), *Trust Management VI* (Vol. 374, pp. 1–16). Springer Boston. Retrieved from http://www.springerlink.com/content/0744828h7136t146/abstract/

Jøsang, A., Ažderska, T., & Marsh, S. (2012). Trust Transitivity and Conditional Belief Reasoning. In T. Dimitrakos, R. Moona, D. Patel, & D. McKnight (Eds.), *Trust Management VI* (Vol. 374, pp. 68–83). Springer Boston. Retrieved from http://www.springerlink.com/content/y7467n0n1177753u/abstract/

Biography

Tanja Pavleska was born on 6. 9. 1985 in Stip, Macedonia. She lived and finished high school in Berovo, Macedonia. She graduated System Engineering and Automation at the Faculty of Electrical Engineering and Information Technologies, University Ss. Cyril and Methodius – Skopje, Macedonia in 2008. After graduation she worked in NIK systems – Skopje, in the department for industrial automation.

In 2009 she joined the Laboratory for Open Systems and Networks at Jožef Stefan Institute and started the doctoral studies at the Jožef Štefan International Postgraduate School in Ljubljana. She has been working as a researcher in the field of Internet trust, security and privacy. The major part of her research is analysing the micro and macro-behavioural phenomena in computational trust and reputation systems, in particular their causal relationship to user bias.