The Logic of Trust

William Thomas Harwood

Ph. D.

University of York
Computer Science

August 2012

Abstract

This thesis addresses two problems of trust:

1. *Knowledge on Trust*: If we are provided with information by a variety of individuals, whom we trust to different degrees, what is the best overall theory we can form from the information we are given?

2. *Social Trust*: If one does not have direct experience of an individual, how can one establish an initial degree of trust through the offices of society?

It addresses the first problem by developing a formal, mathematical and computational, model of Bonjour's Coherence Theory of Knowledge, and the second by adapting abstract argumentation theory to reason about networks of relationships of trust and distrust between individuals. In developing the latter it develops a notion of generalised argumentation systems, giving their semantics via the Galois Connections induced by binary relations, and provides a general scheme of evaluation of these systems based on propositional model finding.

Throughout, some effort is made to set the work in the context of both theories of trust and of the day-to-day trust situations that one encounters in everyday life.

Contents

List of Tables

List of Figures

Preface

Much of what we do in life is governed by one activity – making sense of things. It is something we do every day as part of all sorts of activities, be it dealing with the round of social life or finding explanations in the physical and social sciences. We collect pieces of information from disparate sources, we supplement this information by filling in missing pieces from experience (or prejudice) and create a coherent account of a piece of life. We do not know if this account is *true*. It is not knowledge, but it is belief. It is our best account of what *is*, given the information available to us and our perspective upon the world. This is the sense of logic that is used in the title. Logic, in its most general form, is the process of making sense of things.

Our lives as social animals are bound together by trust, and much of the quality and success of our lives is governed by how well we answer the question of the still small voice inside us, that from time to time asks: *does it make sense to trust this person?* And answering it is a process of making sense. Most of the time the answer will not be incontrovertible. The decision will be a matter of judgement, based on reason, based on logic. This dissertation is my attempt to explore how such notions can be formalised, using the tools of discrete mathematics, familiar enough to a computer scientist, but still rarely found in the disciplines of philosophy and the social sciences that have been the sources of the primary actors in this field.

Acknowledgements

First and foremost I would like to thank Helen, my wife, who accepted the insane idea that I should give up a perfectly good industrial career and undertake a Ph. D. Without her continuing support and encouragement this work would not have been possible.

Next I must thank Professor John McDermid, who, on hearing I wanted to undertake a Ph. D. found funding under the ITA programme and a place within his department at York. And of course I must thank the ITA programme itself for providing both the funding and a research community within which to work[1].

Naturally I must extend thanks to those at York who helped and encouraged me throughout, and in particular I would like to thank my supervisors Professor John Clark and Dr. Jeremy Jacob, and my internal assessor, Professor Jim Woodcock, for their assistance throughout.

Finally I would like to thank all those people with whom, over the years, I have had stimulating conversations on the subjects of computer security, trust and trustworthiness. Necessarily there is something of your ideas in this work. For which I thank you all.

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

Author’s Declaration

The material herein is original work by the author. Some material in this dissertation is in part based on previously published work at peer reviewed workshops and conferences. The material in Chapter 5 is a revised and extended version of the papers “Boolean Coherence: Does it make Sense?" [50], and “Boolean Coherence and the ACH Method” [53]. An early version of the material in Chapter 5 can be found in the paper “Networks of Trust and Distrust: Towards a Logical Reputation System” [51] and the essential philosophy and outlook on the problem of trust on the Internet that pervades the dissertation can be found in the extended workshop abstract “Conceptualising Internet Trust” [52]. These papers are provided as Appendices D, E, F and G to this dissertation.

1. Introduction

1.1. Why Study Trust?

Today nearly everyone seems to be talking about "trust".

Bernard Barber, 1983

The opening sentence of Bernard Barber's "The Logic and Limits of Trust" could have been written today.

Consider the role of trust in the following three recent news stories.

The "banking crisis" escalated from an issue about the value of investments based on sub-prime mortgages to a global banking crisis as banks lost trust in one another over disclosing risk in their portfolios.

The UK Liberal-Democrat party leader, Nick Clegg, has possibly done irreparable damage to his party by, as many people see it, breaking trust over the matter of a pledge about tuition fees.

In the WikiLeaks case, a trusted individual, American soldier, private Bradley Manning, broke trust and disclosed a large quantity of data. This leak was, at best, embarrassing to many, and, at worse, at least for some, placed life or freedom in jeopardy[1].

These stories engage our interest not only because of the consequences they have for us but also because they underline the role of trust in our society.

Trust binds us together. It allows us to live together in relative freedom, work together for common goals and delegate power, authority and action, to others. It allows us to conduct business-as-usual without taking undue risk. The failure of trust means that business-as-usual ceases to be an option.

However there are other stories that are equally about trust that barely graze our collective consciousness. These stories have less impact because they are bound up with the technical world of the Internet. Such stories are often reported as technical, business or personal failures rather than as anything indicative of an underlying flaw.

Here are five Internet related news stories from recent years. They raise questions about the nature of trust in the context of the Internet.

In April 2010 iTunes accounts were hacked and the stolen account details sold on the Chinese market. The purchasers of these accounts could then buy products at the iTunes store and download them to their machines with the charge going to the real account owners. An examination of the mailing list `https://discussions.apple.com` indicates this problem has not been fixed and is still ongoing.

[1] This case also illustrates that although trust and security are related they are different notions. Bradley Manning was using a technical secure computer system at a secure location. Unfortunately, at least for the US government, Bradley Manning was, in this context, untrustworthy.

Such incidents are not isolated. In April 2011 Sony Playstation accounts were hacked and the personal details of 77 million players were lost and the Playstation network taken down for a week. Two weeks later in May 2011 a second hack lost details of a further 25 million, and in October of 2011 details of 93 thousand accounts were lost (on this occasion there was no significant network outage and the accounts were quickly locked to prevent access).

In June 2011 DigiNotar, a Dutch based Certificate Authority, was hacked. As a result, in July 2011, DigiNotar signed falsified certificates undermining SSL, which, for many, is the backbone of Internet security.

Each of these events had real financial, and other, consequences. Apple bears the cost of the hacked accounts both in financial terms and loss of confidence in its services and level of customer care[2]. Sony has suffered both from lost confidence by its users and from the loss of confidence by investors with a 4% drop in share price being directly attributed to the earlier hacks. DigiNotar has declared bankruptcy resulting from the general loss of confidence in its service[3]. Meanwhile, Google, Adobe, Mozilla and Apple have been forced to create updates to purge DigiNotar certificates from their products and Microsoft has delayed a major product update to accommodate the required changes. But this is far from the real cost. Falsified certificates include certificates for Microsoft, Google, wordpress, Facebook, Twitter, Skype, Thawte Root, VeriSign Root, GlobalSign Root, CyberTrust Root, DigiSign Root, and others. For most of these we do not know what, if any, use has been made of them. However, we do know the falsified certificates for Google were used for man-in-the-middle attacks against 300,000 Iranian users[4]. The consequences for these users are unknown. Finally DigiNotar also issued certificates for Dutch government departments meaning that at least some governmental material may be vulnerable, including material shared with Holland by other nations.

The next two stories are about relationships between individuals. And each has tragic consequences.

In March 2010 sentence was passed in two unrelated cases. Peter Chapman, aged 33, was found guilty of the rape and murder of Ashleigh Hall, aged 17. Ashleigh had "met" Peter when Peter posed as 17 year old Peter Cartwright on Facebook. Ashleigh arranged to meet with Peter Cartwright and arranged by texting to be met by Peter Cartwright's father in a car. Posing as Peter Cartwright's farther, Chapman met Ashleigh and drove her to Thorpe Larches near Sedgefield and attacked her. Peter Chapman was found guilty on multiple charges and sentenced to 35 years imprisonment. The second crime was the murder of Camille Mathurasingh by her ex-boyfriend Paul Bristol. Paul's motive was jealousy. After seeing a picture of Camille with another man on Facebook. Paul flew 4000 miles to carry out the killing.

[2] Perhaps what is most surprising here is how well Apple have managed to stop this becoming a larger story.

[3] Although it should be noted DigiNotar is only the most extreme example in a year of Certificate Authorities being hacked.

[4] i.e. users thought they were talking securely to e.g. gmail servers but were in fact talking to an intermediary that could read everything before passing it on to Google.

Paul Bristol was sentenced to life imprisonment[5].

These stories cause us to question trust on the Internet. Can we trust services on the Internet? Can service providers keep our details safe? Can we really rely on security mechanisms to make the Internet safe? Can we trust people we "meet" on the Internet? What assumptions can we make about identities? What assumptions can we make about information we post on the Internet and where that information ends up? What promises are being made to us when we use the Internet? Who makes these promises? And most importantly: What is a rational basis for believing such promises? That is: What is a rational basis for trust?

The problem of finding a rational basis for trust is neither new nor unique to the Internet. As societies evolve, as new business practices and technologies occur, the situations over which we must exercise trust change and we must discover how trust can be underwritten in these new situations. We may think of these as *modes*[6] of trust, Society has already learned to cope with many forms of complex trust relationships: A promise made *now* will be fulfilled at some point in *the future*. A promise made *here* will be fulfilled *elsewhere*. A promise made by *me* will be fulfilled by *another*. A promise made in a *letter*, i.e. made without you and me being in the same place, at the same time, will bind me to what is promised. And so on. Each new modality introducing its own grounds for uncertainty and raising the question "What now underwrites this promise?".

Our challenge is to examine the modalities of trust created in Internet usage and to give an account of what underwrites them.

1.2. Why Study Trust as a Computer Scientist?

As computer scientists we strive to understand how to build systems that are fit for purpose, function correctly and, in the face of unexpected eventualities, fail gracefully and recover cleanly. Today that means understanding the role that computer systems play in the mediation of everyday activities that involve data management and communication, and the interplay between the concepts of security and trust.

Security is about enclosures. It is about locks and gates and fences. Trust is about when you need them and why they might work in the social environment. As computer scientists we need not only to know how to build the enclosures but also to know when they are required and how they can be made to work in the social environment in which they are embedded. And, amongst many other things, this requires us to understand how trust arises and propagates in society.

But there is another way of reading the question above, and that is: "What does a computer scientist bring to the study of trust?". The answer to this is more difficult and, perhaps, more personal. For my own part, it is that a computer scientist brings a particular *outlook* and *set of tools* that differ from those of say, a sociologist or a physicist. Each discipline sees the world in certain terms. For example, the sociologists sees the world as *social processes* whereas the physicist sees the world as *physical processes*. In contrast, the computer scientist sees the world in terms of

[5] Normally 22 years followed by release under license.

[6] mode: A way or manner in which something occurs or is experienced, expressed or done. [OED]

information processes and the tools that s/he uses are for describing and analysing information processes. This difference in perspective leads both to the formulation of different questions and the provision of a different style of answer.

1.3. Overview

In daily life, in order to bring about our goals, we must rely on others. We rely on them to provide us with information and we rely on them to carry out actions. We do this for a variety of reasons, including laziness on our part, economic benefits that might accrue from specialisation, or because what we wish to achieve is beyond our sphere of control or competence. In relying on others we open ourselves to uncertainty, and we need to manage that uncertainty or suffer the debilitating consequence of being unable to act.

When uncertainty exists, society provides many mechanisms for managing the uncertainty, such as, contracts, insurance and hedging. But much of our daily life is conducted relying on others without the use of such mechanisms.

In these circumstances we say "we take it on trust" that so-and-so will act in a particular manner. Indeed, if we do not do this, life becomes unimaginably complex, as we attempt to establish the network of contracts, insurance and hedging required to replace the informal notion of taking things on trust.

But what is trust?[7] This thesis starts from the idea that interaction in society is regulated by the making and keeping of promises, which may be given implicitly or explicitly. It then defines trust to be the belief that people will keep their promises.

The problem then addressed is: What is sufficient reason to believe that these promises will be kept?

Crudely, there are two mainstream theories of trust. What we may term the classical theory of trust argues that trust arises from "thick relationships" which develop between individuals over time where individuals have an opportunity to build up a picture of the promise-keeping behaviour of others.

The other approach is Swift Trust introduced by Meyerson, Weick and Kramer to accommodate situations in which the classical conditions for trust do not exist. In their explanation, trust is founded jointly upon the roles inhabited by participants in an interaction, and upon the perception by participants that they share certain social categories with one another. Thus, for example, in an interaction at a doctors surgery, I extend trust to the doctor because he is in the role of a doctor, at least as far as medical discussions go. I might also extend trust because, for example, the doctor shares some category with me, such as coming from the same town or belonging to the same sports club. Role and category differing in that role confers trust in some specific competence and intent that the doctor should possess (summed up as medical expertise and the hippocratic oath) and category confers some more general notion of trust in some overall intent (e.g. I believe people from my town stick together and treat one another properly).

[7] Chapter 2 addresses this topic in some detail. Here we present enough of a discussion to set out the main direction of this thesis.

However both theories ignore a fundamental aspect of being a social animal with the gift of language. It is possible to learn things from the experience of others *by being told*. In dealing with other people we do not need to use either classical trust or swift trust, we may simply be told about an individual, by someone whose opinion in the matter, we already trust.

This thesis considers this, often overlooked, source of trust, the *embeddedness* of individuals in social information networks that provide them with information about other individuals and circumstances with which they have no direct contact. The term embeddedness here is used to capture the idea of established and lasting relationships as opposed to those which are temporary and ad hoc. Because we are embedded in a network of relationships we can make judgements about the trustworthiness of the individuals with whom we are directly connected. And those individuals can make judgements about the individuals to which they are directly connected. And so on.

Using the social networks in which we are embedded we can synthesise plausible views of the circumstances and trustworthiness of socially remote individuals.

This thesis considers two ways in which individuals can use their embeddedness in their social information network:

1. *Knowledge on Trust*: If we are provided information by a variety of individuals, whom we trust to different degrees, what is the best overall theory we can form from the information we are given?

2. *Social Trust*: If one does not have direct experience of an individual how can one establish an initial degree of trust through the offices of society?[8]

The goal is to answer these questions both mathematically and computationally. That is, to give precise meaning to the concepts involved and to propose a way in which the answers to questions might be computed in particular circumstances.

1.4. Mathematical Approach

In order to gain mathematical and computational leverage, other researchers (for example, Stephen Marsh in "Formalising Trust as a Computational Concept" [68], and Audun Jøsang and Stephane Lo Presti in "Analysing the Relationship between Risk and Trust" [59]) have started their enterprise by assuming some numeric framework for representing degrees of trust and degrees of uncertainty[9]. This

[8] This is not the same as pooling information through the reputation systems commonly discussed in computing circles. The common assumption of rater anonymity is missing. Anonymity has its advantages but has two serious disadvantages. First the rating is *without consequence* to the rater. This means that the rater is free from pressures that might bias their view but also free from pressures that otherwise might keep them honest. Secondly it is not possible to judge the relevance of the rater's experience. Rating from novice and expert, genius and idiot weigh equally. Here, rather than thinking of assessing reputation through anonymous, aggregated, sources, we consider assessing reputation through known, distinct, sources. The effect of this change is that we may consider the reputation of the source as well as the report by the source, in arriving at a conclusion.

[9] And indeed this may be the most appropriate vehicle for many applications.

raises the question, Is such a representation always necessary, or can we make useful progress by choosing other means? This thesis sets out to use the tools of logic and discrete mathematics to see what progress may be possible without the assumption of numeric representations and seeks other structures that might align with informal social processes of trust that we might recognise.

1.4.1. Knowledge on Trust

The problem of Knowledge on Trust is approached by creating a formal model of Bonjour's Theory of Empirical Knowledge. Information sources are modelled as providers of boolean assertions about the state of the world. The goal is to determine a best, or most plausible, set of assertions that can be obtained by combining the assertions from the sources. However, sources may contradict one another. Sources come equipped with a preference ordering that indicates that some sources are preferred over others. Inconsistencies are resolved by discarding information from the least preferred sources contributing to the inconsistency, and information is only discarded to resolve inconsistency.

The model is formulated in terms of forming maximally consistent sets of propositions under the constraints induced by the preference ordering. The approach bears many similarities to approaches taken to both paraconsistent logic and non-monotonic logic.

1.4.2. Social Trust

Modelling of Social Trust is approached through Argumentation Theory and this presents a choice in the possible mathematical tools that we might use to develop argumentation theory and the computation techniques we might adopt to calculate social trust.

The theory may be developed as graph theory, as an application of the theory of binary relations or, more novelly, as an application of Galois Connections induced by binary relations[10]. Here the latter approach is chosen since it provides an elegant tool for developing a uniform and systematic account of Argumentation Theory as used for modelling Social Trust[11]. On the computational front an equally novel account is given of calculating the solution to Argumentation Systems that model instances of Social Trust. A general approach is presented of compiling quite general argumentation systems into Boolean Networks which can be "solved" by use of a model checker[12] (i.e. all models satisfying the boolean constraints are found).

The compilation process, and the connection between the formal models of Knowledge on Trust and Social Trust, rely on the proofs of a number of proposi-

[10] Standard references for Graph theory include Berge's "Graphs and Hypergraphs" [10] and Haray's "Graph Theory" [46]. Standard references for the theory of binary relations are harder to find but Schmidt and Ströhlein is an excellent source, as is Schmidt's relatively newly published "Relational Mathematics" [101]. Detailed references for Galois Connections are given later in the chapter.

[11] And seems to be a novel approach to argumentation theory in general.

[12] For this thesis MACE4 was used for this purpose.

tions. The proofs are given as (mainly) algebraic manipulations of formula about Galois Connections.

1.4.3. Connecting Knowledge on Trust and Social Trust

There is notable similarity between the formal process of assessing Knowledge on Trust and assessing Social Trust. The connection between the two is that they are both consistency based processes and we can simulate some aspects of argumentation consistency within the Bonjour coherence model. A short investigation is made of this connection in Chapter 6.

1.5. Outline of Thesis

Chapter 2 surveys the philosophy and sociology of trust and introduces an analysis of trust as the rational belief that another agent will keep its promises. The chapter then sets out the consequent fundamental questions of trust that arise from this analysis, that is:

- How to combine information from sources of different degrees of trust to arrive at a most plausible theory, or explanation, based on those sources.
- How to decide who, or what, to trust.

Chapter 3 briefly sets out the technical tools of binary relations and the Galois Connections induced by a binary relation.

Chapter 4 sets a framework for addressing the first problem introduced in Chapter 2, based on a formalisation of Bonjour's Theory of Empirical Knowledge. Here the argument is that reasoning from sources of information of different degrees of trustworthiness is the same problem as reasoning from any form of uncertain information. The main obstacle to applying standard probability techniques in this area is that of not having an adequate basis for assigning probabilities to "trust events". As a result this chapter takes an alternative approach of working with preferences expressed over information sources and develops a logical, as opposed to probabilistic, analysis of the notion of most plausible theory given the preference structure on the information sources.

Chapter 5 sets out an approach to the second of the problems introduced in Chapter 2, that is who or what to trust. This analysis is developed using argumentation theory as a tool to define a logical notion of reputation. In the process of developing the framework used, a substantial analysis of argumentation systems is developed.

Chapter 6 shows that formal models developed in Chapters 4 and 5 are linked by the notion that there is at least a partial simulation of one model by the other.

Chapter 7 briefly discusses a number of themes that arise out of the technical analysis of trust offered in earlier chapters.

Chapter 8 draws the thesis to a close by discussing the significance of the models of trust and privacy.

The Appendices As usual appendices have been used for additional materials, which although interesting or useful, would disrupt the flow of the dissertation.

1.6. Research and Published Papers

My research has progressed by weaving backwards and forwards between social science and computer science. Part of this process has been trying out ideas on different audiences concerned with social science, computer science and logic. The workshop papers that resulted from this reflect my thinking about the topics of this dissertation at particular points in time (these papers are included as appendices D-G). On the whole, unlike many dissertations, the chapters are not amplifications of the papers. Rather the chapters represent the result of feedback, reflection and further research on the topics. That is to say, this research has evolved, progressing by the age old process of *performance, feedback* and *revision*. Or, as Babba Brinkman puts it more lyrically:

> *And sometimes people ask:"How does your show get written?"*
> *Like this: performance, feedback, revision.*
> *And how do I generally develop my lyricism?*
> *Like this: performance, feedback, revision.*
> *And how do human beings ever learn to do anything?*
> *Like this: performance, feedback, revision.*
> *And evolution is really an algorithm that goes*
> *Like this: performance, feedback, revision*

The Rap Guide to Evolution

Lyrics by: Baba Brinkman, 2011

1.7. Contribution

The central contribution of this dissertation is the analysis of trust as arising from social embeddedness. In particular, it offers a theory of trust and trustworthiness that is based on an informal theory of promises and promise keeping[13]. This informal theory leads to considering two specific questions in relation to trust and trustworthiness. How do we evaluate Knowledge-on-Trust? And: How do we evaluate Social-Trust?

[13] The formalisation of this theory is proposed as an area for further work in the conclusions.

These questions are explored using formal modelling. In the exploration a number of technical contributions are made to the Coherence Theory of Knowledge and to Argumentation theory.

The contribution to the Coherence Theory of Knowledge arises in Chapter 4 in the creation of a formal, mathematical and computational, model of Bonjour's Coherence theory and the application of this formalisation to a number of example situations requiring reasoning from information gained from sources of known relative trustworthiness to plausible conclusions.

In the case of argumentation theory, the technical contribution is the development in Chapter 5 of:

- A Galois Connection approach to the argumentation theory,
- A generalised framework for argumentation systems extending both Dung's argumentation systems and bipolar argumentation systems,
- A computational model of argumentation based on boolean model satisfaction.

This computational model has resulted in a new notion of sensitivity analysis of argumentation systems which leads to a natural notion of the degree to which the one argument affects another in forming a position (i.e. a mutually acceptable collection of arguments). This analysis is discussed in the penultimate chapter (Chapter 7) as one of the areas for further work.

Along the way new connections have been made between disciplines. In particular, the interaction between theories of trust and trustworthiness and the world of computer security access control mechanisms, and the connection between argumentation theory and the social sciences notions of Catnets, and Structural Balance Theory, are discussed.

Finally, although in no way central to this dissertation, a brief outline is given in appendix C of a novel approach to information flow security based on Galois Connections arising from binary relations.

1.8. The Personal Journey

The concepts I hold about trust and trustworthiness at the end of this piece of research are quite different from those I held at the beginning. Originally I conflated the concepts of trustworthiness, reliability and risk and confused the notion of a logic of trust with a logic of risk. My view now, as I hope is evidenced in this dissertation, is that the concepts are quite distinct, even though we may, in informal speech, tend to use the *words* interchangeably. Trust requires a promise made, reliability requires an expectation, but not a promise, and risk is a measure of uncertainty that trust is intended to offset. These points are made more elaborately in Chapter 2. If one accepts such distinctions, as I do, but others may not, it becomes natural to look at trust as a qualitative rather than quantitative phenomenon. This dissertation is an attempt to work through such a view.

2. Theories of Trust

In which we discuss:

- *Theories of Trust.*
- *How Trust is related to notions of co-operation and risk.*
- *Some problems of modelling trust as subjective probability.*
- *How we can view trust as beliefs about keeping promises.*

2.1. Introduction

We have limited control over the world in which we live. To achieve many desirable things we must rely on the actions of others. But when is such a reliance a reasonable, or indeed a rational, choice? What justifies our belief that another will indeed act in a way beneficial to ourselves?

Over the last 30 years there has been an increasing interest in the study of trust. This interest has arisen from the concern that in many western societies trust is in decline, both in terms of generalized trust in government and institutions and, in specific trust between individuals (see, for example, Barber [6]). Much of the discussion of this trend focusses on the shift in structure of society from community based interaction, in which, historically, individuals interacted with small, fixed, communities over most of their lives to network based interaction, in which individuals interact, directly and indirectly, with large numbers of others through ever-changing networks of personal relationships and commerce [49]. The main contention of the social theorists working in this field is that this change in society weakens the generalized reciprocity that exists in small face-to-face communities and replaces it by the specific reciprocity of direct interaction in the network and *systems trust* (see Luhmann [66, page 68]) which is the trust "in the ability of the systems to maintain conditions or performance".

It is useful to categorize the extensive literature which has developed into *explanatory* theories and *utilitarian* theories. Explanatory theories explain the role or function of trust in societies and how the notion of trust changes as the structure of societies change. In particular most such theories are concerned with the notion of trust in *modern* societies and how this notion acts as a cohesive force to hold societies together against the internal dissipative pressure of modern life. Utilitarian theories are concerned with the central problem that each of us face as individuals, which is, how to make the decision of when, and when not, to trust and what are adequate conditions for this decision to be considered reasonable or rational?

The focus of this dissertation is a utilitarian theory of trust. To understand the context of such a theory, it is necessary to examine the main themes of explanatory theories of trust but unnecessary to pursue them in great detail. To this end we briefly summarize the main themes of the explanatory theories of trust in modern societies and then discuss utilitarian theories of trust.

The plan of this chapter is to start in section 2.2 with a working definition of trust to orient the discussion. This is followed in section 2.3 by a discussion of trust's near relatives. Section 2.4, Trust and Society, gives a brief discussion of the explanatory theories of trust and some views of the role of trust in society. Section 2.5 then takes up utilitarian theories of trust. Section 2.7 takes up the issue of the role of forgiveness in maintaining trust relations. Section 2.8 summarizes the concepts covered before the final section 2.9 briefly introduces a synthesised theory of trust and trustworthiness that seems appropriate to our studies.

2.2. An Orientating Definition of Trust

To anchor the discussion, we give a working definition of trust, at least on a temporary basis. Barber[6] describes trust as generalized expectancy of good behaviour on the part of others over whom we have no control. Kydd [64, page 6], in the context of international negotiations, provides a straightforward operational definition "trust is a belief that the other side prefers mutual co-operation to exploiting one's own co-operation, while mistrust is a belief that the other side prefers exploiting one's co-operation to returning it. In other words, to be trustworthy, with respect to a certain person in a certain context, is to prefer to return their co-operation rather than exploit it." This view of trust in terms of beliefs or dispositions neatly deals with several thorny issues. First it explicitly defines the relationship between trust and trustworthiness. Second, by being disposition based it defines trust in situations where meaningful action is not possible (what does it mean to say "I trust the government" mid-term, when I have no recourse to a meaningful action such as voting? c.f. Hardin's discussion [49]). Third, it clearly says that trust is contextual i.e. applies to "a certain person in a certain context". Fourthly, and finally, it says that trust is a determining factor between the expectancy of co-operation and expectancy of exploitation.

Hardin[48] adds that the trust is about something in particular. That is, we trust people for particular things, e.g. for being knowledgeable about some particular discipline, or for returning loans promptly, rather than some blanket trust in which we trust them in whatever they do or say. As Hardin puts it "trust is a three part relation A trusts B for C".

Many regard trust as originating as a component of generalized exchange [28, pages 97-14]. That is, A and B wish to exchange goods regarded of equal value. Ideally they would exchange goods simultaneously, and in the same place, so that neither party may cheat in the arrangement. However, simultaneous exchange is not always possible. Generally, exchanges may require the two parts of the transaction to be separated in time or space (or both). Although it is often implied that such distributed exchange is a modern phenomenon it can be found in antiquity. For example, at the time of the Crusades the Knights Templar offered one of the first international banking services in which a Knight could deposit wealth with the Knights Templar in his home land and draw on that wealth in the Holy Land. Undoubtedly this was a distributed transaction in time and space involving a large component of trust. The real difference between traditional and modern society is that of the extent of distributed transactions. Once they were the exception, today, at least for many of us, they are the rule. Trust's role in distributed exchange is to remove the uncertainties and doubts that one party may have about the intentions of fulfilment of the other. And as Kydd's phrasing makes clear, the exchange may not be of material goods but of intangibles such as access to land, rights and permissions to act in certain ways, intellectual property, etc.

What then are the components of a trust based transaction, or generalized exchange? The key elements are set out by Coleman[28] as:

- The trustor cannot achieve his ends, or cannot achieve his ends economically, without the collaboration of the trustee.
- The trustee commits to actions on behalf of, or in the interest of, the trustor (let us call these the trustee's obligations), but the trustor has no control over the trustee's actions and so is subject to uncertainty or risk.
- If the trustee fulfills her obligations then the trustor will be better off than if he had not trusted the trustee. If the trustee fails to fulfill her obligations, the trustor will be worse off than if he had not trusted the trustee.
- Generally the trustor places resources at the disposal of the trustee but has no control over what the trustee does with these resources. So the trustee may use those resources for her own benefit, for the benefit of the trustor, or for the benefit of both.
- As a result of the trust the trustee is in a position to do something that she could not otherwise do.
- Finally, generally, there is a time lag[1] between the commitment to the transaction by the trustor and the fulfilment by the trustee.

With these notions of trust in mind we will quickly compare trust with some of its near relatives and then examine some explanatory theories of trust.

2.3. Trust's Near Relatives

Trust is part of a network of mechanisms that have developed to handle social interactions that are extended in time and space. Other parts of this network include risk management, insurance, hedging, distrust, contract, statutory obligations, power, control, authority, toleration and legitimacy. Each offers a way of reducing uncertainties present in particular forms of social transaction either when used individually or when used in concert.

When we cooperate with others to achieve some goal, we introduce uncertainty in that, at least potentially, we make ourselves vulnerable to the possibility of

[1] We may add also "and/or spatial separation"

exploitative behaviour on the part of others. When, then, may we rationally rely on others to cooperate rather than exploit? There are many possibilities, indeed not all cooperative behaviour requires a social mechanism to support it. For example, if the other parties involved in the co-operation have no choice in their actions: their options are so limited that their action inevitably benefits us. Alternatively they may be acting in their own best interest and the benefit to ourselves is an inevitable consequence. We may explain the perception of good behaviour of others simply in terms of the predictability of the situation. It is in situations where others have real choice of options, some of which benefit ourselves and some of which do not, that we need to start to appeal to some social mechanism to support our own goals.

Some mechanisms, such as *insurance*, reduce uncertainty by spreading risk over many, uncorrelated, transactions. Failure of some transactions is compensated for by success in others and an acceptable average outcome is obtained. In contrast *hedging*[3] handles uncertainty by spreading risk over correlated events so that a negative outcome in one area correlates to a positive outcome in another. Generally both insurance and hedging have an associated cost and trade an overall lower expected outcome for a safer outcome. But the closer relatives to trust manage the uncertainty in a single transaction by filling in the missing information that creates uncertainty.

To understand a little more what trust and its close relatives achieve, consider two alternative games based on the well-known scenario of *Prisoner's Dilemma*. The first game is Prisoner's Dilemma. Two prisoners, A and B, are being interviewed in separate rooms. Each is offered the same deal. If they betray the other prisoner then they will receive no sentence (0 years) but the other will receive a heavy sentence (3 years). If neither betrays the other then they can expect a modest sentence of 1 year but if both betray the other they will each receive a 2 year sentence (see figures 2.1). The prisoners have a clear choice of cooperating with one another or betraying one another. There is a built in bias for each prisoner to betray the other in this game because each prisoner prefers to meet co-operation with betrayal.

		B	
		Cooperate	Betray
A	Cooperate	-1,-1	-3,0
	Betray	0,-3	-2,-2

Table 2.1.: Prisoner's Dilemma

Now consider an alternative game, the *Assurance Game*. The scenario is the same as the Prisoner's Dilemma but the payoff matrix is different. (see figure 2.2). In the assurance game the bias is to return co-operation with co-operation. This is not sufficient to guarantee co-operation since each prisoner needs to believe that the other prisoner will cooperate but, if each prisoner does so believe, then co-operation will occur.

		B	
		Cooperate	Betray
A	Cooperate	0,0	-3,-1
	Betray	-1,-3	-2,-2

Table 2.2.: Assurance Game

The Assurance Game is an instance of a more general game known as the *Stag Hunt*. The difference between the two games is that whereas the Assurance Game is a two person game, the Stag Hunt is an n-person game. The game is a vehicle for discussing the Social Contract and motivated by the following passage from Rousseau's "Discourse on Inequality" [95]:

> *"If it was a matter of hunting a deer, everyone well realised that he must remain faithful to his post; but if a hare happen to pass within reach of one of them, we cannot doubt that he would have gone off in pursuit of it without scruple."*

Since the Stag Hunt is most naturally phrased in terms of positive rewards (the gain of a portion of a Stag or a Hare) and losses (wasted effort) we will rephrase the game matrix used for the assurance game by adding a constant of 2 to all entries:

		Others	
		Cooperate	Betray
A	Cooperate	2	-1
	Betray	1	0

Table 2.3.: Stag Hunt

In this matrix the payoffs are shown for player A with the interpretation of the action by others as, if *all* of the others co-operate and if *any* of the others betray. Of course, the payoff matrix is the same for all other players. In this game if everyone co-operates then the maximum reward is obtained by all, but if one person betrays the enterprise then the betrayer gains a hare and all other players have wasted their time. The action of players depends on what they believe about other players and the likelihood of a successful outcome to a stag hunt. Again, as with the Assurance Game, if all players believe in the co-operation of other players the rational outcome is co-operation.

The different social mechanisms can each be seen as a means of providing the missing belief, or knowledge, that the other will cooperate. Some mechanisms arise from "the general good" of social and legal (statuary) pressures or requirements. But others need to be explicitly applied to a situation such as the creation of a contract or the placing of trust in an individual or institution.

One interesting observation made by Brian Skryms is that iterating the Prisoner's Dilemma, with the same participants playing the game over and over again, with no fixed end in sight, transforms the game to an instance of the Stag Hunt. Or as he says:

> *The shadow of the future has not solved the problem of co-operation in the prisoner's dilemma; it has transformed it into the problem of co-operation in the stag hunt.*

That is, the expectation of future interaction and reward for continuing co-operation changes the game we are playing.

A further insight can be gained from Hollis's *centipede games*[2] which he introduces in [56] with the *penny pinching* game and further elaborates the analysis in his monograph "Trust within Reason"[57]. Penny pinching is a game of two players. A stack of pennies is placed on a table and players take turns at removing pennies. The rules allow a player to either take one penny or two pennies in a turn. If there are no more pennies the game ends. If a player takes a single penny then the other takes a turn. If a player takes two pennies then the game ends. Players keep the pennies they have taken. Consider a game with 8 pennies and two players Alice (A) and Bob (B). Alice starts the game. What strategy should the players adopt? If we examine the game tree for penny pinching we see it has a peculiar structure:

$$\begin{tikzcd}
(0,0) \arrow[r, "A=1"] \arrow[d, "A=2"'] & (1,0) \arrow[r, "B=1"] \arrow[d, "B=2"'] & (1,1) \arrow[r, "A=1"] \arrow[d, "A=2"'] & (2,1) \arrow[r, "B=1"] \arrow[d, "B=2"'] & {} \\
(2,0) & (1,2) & (3,1) & (2,3) &
\end{tikzcd}$$

$$\begin{tikzcd}
(2,2) \arrow[r, "A=1"] \arrow[d, "A=2"'] & (3,2) \arrow[r, "B=1"] \arrow[d, "B=2"'] & (3,3) \arrow[r, "A=1"] \arrow[d, "A=2"'] & (4,3) \arrow[r, "B=1"] & (4,4) \\
(4,2) & (3,4) & (5,3) & &
\end{tikzcd}$$

Here the game position is represented as a pair (Alice's number of pennies, Bob's number of pennies) and the moves are: Alice takes 1 penny ($A = 1$), or Alice takes 2 pennies ($A = 2$), and similarly for Bob.

[2] They are called centipede games because of the shape of the game tree

At each turn, except at the very end when the option is not available, players can do better for themselves by taking 2 pennies if they assume the other player will take 2 pennies on the next turn. Ultimately such reasoning results in a backward induction which is typical of finitely iterated games[104], where the first player concludes that the optimal strategy is to maximise the minimum return by taking 2 pennies at the outset. Hollis's point is that this is unreasonable in that one would naively expect that both players have a preference for the game ending at the high end of the tree i.e. at (3,4), (5,3) or (4,4). However it is always in the other player's "local" interest to end a step earlier i.e. if Alice believes that Bob will end at (3,4) then Alice will do better by ending at (4,2). To obtain the higher payoffs, either one player must accept a below optimal outcome and the other player must believe this to be the case, or both players accept a below optimal (for them) but "fair" outcome of (4,4) and both players believe this. That is, some form of trust is necessary to realise outcomes better than (2,0). The problem is: What sort of additional information can possibly lead to the required trust?

2.4. Trust and Society

The German social theorist Luhmann proposes that the primary role of trust in society is the management of complexity in decision making by allowing the possibility of actions in which benefits are deferred and the interval of acting is expanded

"The world is being dissipated into an uncontrollable complexity; … Nevertheless, I have to act here and now. There is only a brief moment of time in which it is possible for me to see what others do …In just that moment only a little complexity can be envisaged and consciously processed, thus only a little gain in rationality is possible.…If I can trust …, I can allow myself forms of co-operation which do not pay off immediately and which are not directly visible as beneficial. If I depend on the fact that others are acting, or failing to act, in harmony with me, I can pursue my own interests more rationally."[66, page 24]

The role of trust, in Luhmann's theory, is to create certainty in the face of uncertainty and thus allow action as opposed to endless indecision. Barber offers us a similar insight. Building on the notion of justified expectation, he says "Trust …has the general function of social ordering, of providing cognitive and moral expectation maps for actors and systems as they continuously interact"[6, page 19]. That is to say, trust makes things predictable in a way that allows us to act.

Misztal elaborates this line of thought by discussing the role of different forms of trust in creating different forms of order that reduce or eliminate various kinds of uncertainty. O'Hara[81] provides a basic summary of Misztal's view in Table 2.4 showing various roles or functions of trust and the underlying mechanism by which these are achieved. Misztal's work is integrative bringing together many themes from across the spectrum of explanatory theories.

Here we see the roles or functions of trust across different forms of interaction. Misztal's own summary[76], Table 2.5, introduces additional notions of the kind of order created in society by a particular kind of trust and the social mechanism that the particular form of trust arises from.

Trust: What it does	How it does it
Makes things predicable	Habit and routine
	Living up to reputation
	Remembering
Brings us together	Family
	Friends
	Society
Helps us work together	Solidarity
	Toleration
	Legitimacy

Table 2.4.: Misztal's Synthesis of Sociological Theories of Trust

Order	Trust	Practice
Stable	Habitus	Habit
		Reputation
		Memory
Cohesive	Passion	Family
		Friends
		Society
Collaborative	Policy	Solidarity
		Toleration
		Legitimacy

Table 2.5.: Trust: Forms and Practice

"Habitus is a system of dispositions 'a past which survives in the present and tends to perpetuate itself into the future by making itself present in practices structured according to its principles' " [76, page 97]. Habitus is acquired by an agent through participation in society and may be seen as the set of implicit *rules* or *conventions* that exists below conscious and rational thought, acquired by the process of habituation, that govern social interaction in a given society. Part of these rules and conventions perpetuate Habitus itself (as reflected in the response to "Why do it this way?" being "Because we have always done it that way"). Habitus gives rise to the predictability and stability that comes from the expectation that people in society will implicitly follow its rules and conventions and thus reduces the uncertainty of interaction.

Trust as Passion is trust arising from the development of the intimate social bonds between people based on shared values and experiences. It is the "internalized trust" that arises from those basic face-to-face interactions in small societies and its role is to hold together groupings within society, providing a basic integrating and cohesive force.

Trust as Policy operates at the conscious dispositional level. That is, trust as policy is built on the set of dispositions that we consciously hold towards individuals and groups. Misztal [76, page 100] quotes Dunn's summary "that trust 'is a more or less consciously chosen policy for handling the freedom of other human agents or agencies'".

Each of these forms of trust gives rise to a specific notion of order in society. These notions of order are not seen as competing but rather as complementary principles countering the dissipative forces pushing society apart.

Perhaps the most significant aspect of this analysis is that different forms of trust may be subliminally learnt or may be matters of conscious policy chosen to impose order on a disordered world. To insist that trust was-one-or-the-other would be to make the error that there is only one possible source of "internal certainty".

2.4.1. Abstract Systems and Anonymous Others

In modern societies we have a need to go beyond interaction with known individuals and extend trust to anonymous others. The chains of relationships can stretch around the globe with little possibility of knowing any but a small number of participants in the overall relationship. Most of the elements of the chain are anonymous to us. How then do we reduce uncertainty? It may be possible to assume that there are sufficient ties to a common society that Habitus and Policy are sufficient grounds for trust but often this is not the case.

Luhmann argues that in modern societies, with their vast scope of possibilities for impersonal interaction, the need to manage complexity gives rise to abstract mechanisms of interaction such as Money and Power. These abstract mechanisms, which Luhmann calls *generalized communication media* substitute for the personal relationships built up in face-to-face societies and their role is to provide *equivalent certainty*[66, page 50] to personal trust. But this then leads to a need for *systems trust* which is the belief in the ability of these abstract mechanisms to guarantee that equivalent certainty. Thus Luhmann sees systems trust as the cost, and necessary underpinning, of the modern, networked society.

This theme is picked up in the work of Giddens[42, page 120]: "Trust in abstract systems is the condition of time-space distanciation[3] and of the large areas of security in day-to-day life which modern institutions offer compared to the traditional world."

It is possible to 'trust' anonymous individuals because interaction with them is through generalized communication media, through generalized systems of exchange, that are themselves supported by systems trust. *We may posit the general goal of any online trust framework is either to reflect the societal trust frameworks built on Habitus, Passion and Policy or to establish a new generalized communication medium to carry the burden of trust. In either case it must be supported by systems trust underwriting the particular mechanism.*

2.5. Individuals

Against this background of the functions and roles of trust in society we now ask: What is trust from the individual point of view? What does it mean for one individual to trust another? For the individual the role of trust is to reduce or remove uncertainty by replacing external certainty with internally generated certainty [66].

A different analysis of trust is provided by examining the reasons behind trustworthy behaviour. Hardin provides a summary of current theories of trust from the viewpoint of the mechanisms underlying trustworthy behaviour. For Hardin trust and trustworthiness are inseparable concepts and he contends that most theories of trust are, in fact, theories of trustworthiness [49]. He divides theories of trustworthiness into three types which we may call, moral, social and relationship theories. Moral theories explain trust as arising from inner moral concerns of individuals. Social theories of trustworthiness contend that trustworthiness arises from the effective possibility of social sanction i.e. an individual is trustworthy in so far as the social norms of behaviour are concerned and trustworthiness is maintained by social sanction for erring from the acceptable norms. Relationship theories of trustworthiness maintain that trustworthiness arises from the desire to maintain a relationship between the trusted and the trusting parties and failure to meet one's trust obligations results in damage to, or termination of, the relationship.

These particular explanations of trustworthiness may be set into a general scheme. If A regards B as trustworthy for C, then A believes that there is a property, ϕ, that B wishes to maintain and, if B undertakes an obligation to A for C, then there is a mechanism whereby ϕ will be damaged if B fails to meet the obligation to A. Placing Hardin's examples into this framework is straightforward. For example if we take moral theories we have the property being maintained as "moral integrity" and the enforcement mechanism may either be an internally imposed mechanism, such as an individual's need to maintain self image and avoid cognitive dissonance, or externally attributed mechanism such as "the wrath of god" at immoral acts[4]. Similarly the social and relationship models may be cast

[3] the stretching of social relations across time and space

[4] There is no commitment in this to the external agency being real or actual. The framework is one

into the framework. However we may also develop other examples that may use other principles or be more specific than these general categories.

For example we may consider A buying from an online store B where C is the fulfilment of the order. A believes that B wishes to maintain profitability. If B undertakes the order, an obligation is created from B to A for C. If B fails to meet this obligation then B's profitability will be damaged via the mechanisms of reputation, the credit card company (via financial sanctions and, ultimately, withdrawal of service) and, ultimately, whatever legal framework that provides customer protection in the domain in which the transaction was enacted. As this example shows there may be multiple enforcement mechanisms in play and trust may arise from their combination rather than the existence of any one of them.

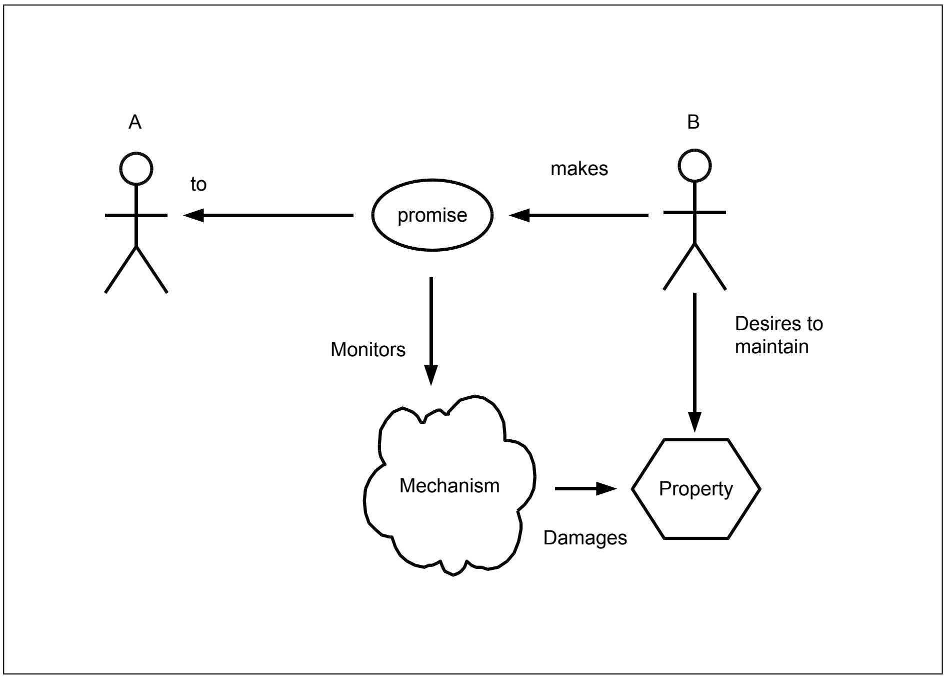

Figure 2.1.: Generalized Model

Hardin's own theory of Encapsulated Interest[48] is a relationship theory in which one individual, B, encapsulates the interests of another individual, A, by making a commitment in regard to A's interests. This obligation is enforced through the desire by B to maintain the relationship with A. However the commitment is not a guarantee to fulfill the goals of A. Rather it is a commitment, by B, to give A's interests suitable regard in taking action. It is possible, for example, that fulfilling A's goal conflicts with other obligations on B, or directly with B's own goals. Here a demonstration that B failed to give adequate consideration to A's goals, or indeed A's belief that this was so, leads to the relationship being damaged. Again this model may be cast into the form above.

of beliefs about beliefs. If I believe that God exists and his wrath will be visited upon me for immoral acts and you believe that I believe this, then you have an adequate basis to trust me.

2.5.1. The Subjective Probability Theory of Trust

Trust as subjective probability has been a popular theme in computer science approaches to trust (see, for example, Jøsang et al. "A survey of trust and reputation systems for online service provision" [60]). The rise of this view of trust owes much to the influence of Gambetta's article "Can We Trust Trust" [40] in which he summarizes the results of a significant workshop on trust[41] by: *"In this volume there is a degree of convergence on the definition of trust which can be summarized as follows: trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action (or independently of his capacity ever to be able to monitor it) and in a context in which it affects his own action (see Dasgupta and Luhmann in particular, this volume)."* However no mathematical detail of this consensus view is presented (although there is a model in Dasgupta's paper that uses the subjective probabilities as a given). Two works that do much to set this right are Coleman in "Foundations of Social Theory"[28] and Marsh, in his 1994 doctoral thesis[68]. Both make good cases for trust as subjective probability. For an extensive review of the literature of the subjective probability approach we refer the reader to Marsh. Coleman's is the simpler presentation, which we will adopt here as a basis for a discussion of the issues around subjective probability models of trust.

Coleman proposes that associated with any transaction involving trust there is a subjective probability p of the trustee fulfilling their trust obligations. The trust situation is described by a gain of G going to the trustor if the trustee fulfills their trust obligations and a loss of L occurs to the trustor if the trustee fails to fulfill the trust obligations. The proposal is then that the trustor should make the decision to trust or not to trust according to:

$$\begin{array}{lll} \text{to trust} & \textbf{if} & \frac{p}{1-p} > \frac{L}{G} \\ \text{neutral} & \textbf{if} & \frac{p}{1-p} = \frac{L}{G} \\ \text{not to trust} & \textbf{if} & \frac{p}{1-p} < \frac{L}{G} \end{array}$$

Or, equivalently in terms of expectations:

$$\begin{array}{lll} \text{to trust} & \textbf{if} & p.G > (1-p).L \\ \text{neutral} & \textbf{if} & p.G = (1-p).L \\ \text{not to trust} & \textbf{if} & p.G < (1-p).L \end{array}$$

Coleman illustrates how this simple model can account for a number of trust situations, provided that the probability and the levels of potential gain and loss are known or can be reasonably estimated. However the model is not without its faults. In many situations we must assume that certain losses are in effect infinite so no amount of effective gain could tempt us into gambling on absurdly low probabilities (does a rational person with a good lifestyle bet everything they have for the remotest of possibilities of winning 10 billion pounds?). The St. Petersburg paradox forces us even further. Consider the game in which I am willing to give you a pound if the toss of a coin comes up heads but if the coin comes up tails I will repeat the bet with double the reward and I will continue to do so until a head occurs. To play this game I request you to pay a fair sum. What is a fair sum? The expected gain of this game is infinite. You should be willing to pay all of whatever you have to play the game. But no rational person would, because the most likely outcome is that a head will occur early in the sequence leaving you with a very small sum compared to what you have given up. This is the St. Petersburg paradox and has been known since the early days of probability theory[11][5]. It is usually countered by the switch from values to utilities i.e. the actual worth of the benefit/loss to an individual. The basic assumption of utilities is that utility tails off as value increases e.g. You may have a high utility for the first million pounds but it goes down rapidly between £1 million and £10 million to the point where increases in return become worth nothing. The effect of this switch is to limit the valuation of the game to something that results from a small number of plays. However in many respects this is still far from satisfactory. Although you may have real utility for £1 million you are not willing to pay £1 million to participate in the game. Your belief is that the game will terminate much earlier since the termination probability is $\frac{1}{2^n}$. So there is only approximately a chance of $\frac{1}{1000}$ that the game will make it 10 steps. Even if the utility based approach is taken, as in Marsh, there is still the problem that the St. Petersburg paradox can be rewritten using utilities rather than values.

If we accept some framework for avoiding these problems we are left with two questions:

- The subjective probability *of what*?
- Where do the probabilities come from?

To say that the subjective probability is a measure of how likely another agent is to carry out a particular action is really to beg the question. It in effect says that it is a measure of trustworthiness without indicating what would actually make another agent trustworthy. Without such a definition we cannot answer how the probabilities would be arrived at. It is difficult to conclude that they can be acquired by observation alone. How can we generalize a run of observations into the future? We can only generalize with respect to some theory of motivation or underlying mechanism[6] but this is precisely what the subjective probability account of trust fails to give us. So although, for example, Coleman says that, when faced with new situations, we may appeal to similar situations we have encountered in the past, Coleman fails to tell us how we would recognise those past situations as similar and applicable.

Perhaps, however, the most significant problem of reducing trust to subjective probabilities is that it ignores the role of trust in reducing uncertainty. If we

[5] Original publication dated 1738. Modern translation by Dr. Louise Sommer, The American University, Washington, D. C., from "Specimen Theoriae Novae de Mensura Sortis," Commentarii Academiae Scientiarum Imperialis Petropolitanae, Tomus V [Papers of the Imperial Academy of Sciences in Petersburg, Vol. V], 1738, pp. 175-192

[6] As Deutsch observes, " *'Trust', involves the notion of motivational relevance as well as the notion of predictability."*

cannot act because of uncertainty then what gain do we have in quantifying the uncertainty? The subjective probability theory of trust gives us a decision criterion for when we may regard it as profitable over repeated events to trust, but does not contribute in any way to reducing uncertainty or removing the paralyzing anxiety that stops us from acting in non-repeated events[7]. The solution to reducing uncertainty is additional information. That is, reasons why an individual may be trustworthy. We will return to this question in the final section of this chapter.

One of the most attractive features of trust as subjective probability is that it provides the starting point for a mathematical theory of trust but it is not the only such possible starting point. If we examine the essay by Luhmann referred to by Gambetta[67], then we may note that Luhmann does not discuss subjective probability. Rather Luhmann discusses *risk* as a concept and we have no particular reason to assume that he is referring to measurable risk, that is probability (subjective or otherwise) as opposed to uncertainty, that is unmeasurable risk. Indeed if we compare the discussion to his earlier work[66] we find risk used in much the same way as uncertainty e.g. *"trust is a solution for specific problems of risk"* with the explanation of risk thus: *"The distinction between confidence and trust depends on our ability to distinguish between dangers and risks, whether remote or a matter of immediate concern. The distinction does not refer to questions of probability or improbability. The point is whether or not the possibility of disappointment depends on your own previous behaviour"*. That is, risk is something that an actor exposes himself to by making a choice, risk is the exposure to potential loss and this exposure generates uncertainty. Trust is the specific solution to that uncertainty.

Whatever one thinks of Luhmann's actual intentions in the matter it is certainly the case that there is an alternative to treatment of risk as unmeasurable uncertainty. Knight[62] proposed a distinction between *risk* which he regards as measurable uncertainty and *true uncertainty* which is not measurable. *"It will appear that a measurable uncertainty, or "risk" proper, as we shall use the term, is so far different from an unmeasurable one that it is not in effect an uncertainty at all. We shall accordingly restrict the term "uncertainty" to cases of the non-quantitative type. It is this "true" uncertainty, and not risk, as has been argued, which forms the basis of a valid theory of profit and accounts for the divergence between actual and theoretical competition."* Knight's contention is that once risk is measurable it becomes manageable by the standard approaches of calculating expectations and spreading exposure over multiple "bets". Uncertainty, on the other hand, is the realm of knowledge and judgment.

The theory of decision making under uncertainty offers alternatives to utility theory. The classic paper "Games Against Nature" by Milnor[75] sets out the standard approaches to decision under uncertainty and creates an axiomatic analysis of each. The following descriptions are adapted from Milnor:

Assume that an individual, the player, is faced with making a choice between a number of options. Each choice results in a different range of possible outcomes. The choices are arranged in a matrix **A** with each choice being a row and the possible outcomes being the row entries. The classical decision procedures are

[7] Presumably, for repeated events at least, one may argue that knowing that one will gain *on average* enables one to act.

then:

Laplace (*Lack of information/Maximum Entropy*) Assume all outcomes are equiprobable. So the value of the choice of the i^{th} row is the average of the outcomes. The player should choose the row that maximises the value.

Wald (*Minimax principle*) Choose the row that minimises the maximum loss. If there is more than one such row choose between them randomly.

Hurwitze Select a constant $0 \leq \alpha \leq 1$ which measures the player's optimism. For each row i (or probability mixture of rows) let a_i denote the smallest component and A_i the largest component. Choose a row for which $\alpha A_i + (1 - \alpha)a_i$ is maximised.

Savage (*Minimax Regret*) Define the (negative) regret matrix $\mathbf{r_{ij}}$ by $r_{ij} = a_{ij} - m_j$ where m_j is the maximum of column j. r_{ij} measures the difference between the actual payoff obtained and the payoff that could have been obtained with perfect information about Nature. Now apply the Wald criteria to the matrix $\mathbf{r_{ij}}$ and choose the row that minimizes the highest regret.

These decision procedures are quite distinct from each other, for example Milnor gives the following matrix with the associated results of the decision procedure.

$$\begin{array}{cl} \begin{bmatrix} 2 & 2 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 4 & 0 & 0 \\ 1 & 3 & 0 & 0 \end{bmatrix} & \begin{array}{l} \text{Laplace} \\ \text{Wald} \\ \text{Hurwitze (for } \alpha > \frac{1}{4}) \\ \text{Savage} \end{array} \end{array}$$

Clearly the procedures are different and each offers an alternative treatment of decision making under uncertainty to that of decision making under risk. It would be possible to build a theory in which different actors had different policies of uncertainty resolution and chose to trust, or not to trust, based on minimising potential loss, internal optimism or regret minimisation. Moreover we might reasonably suppose that, in at least some cases, these are strategies that are actually used by people in the face of real uncertainty. Choice under uncertainty gives a model of the situation when A decides to trust B for C. Specifically A may chose actions with a possible set of outcomes. If A trusts B for C then A is essentially an optimist with respect to his choices and expects a good result. If A distrusts B with respect to C then A expects a bad result. When faced with a number of possible options of who to trust for what, the trustor may apply any of the strategies for deciding who to trust.

Subjective probability does give an account of certain type of trust situations but even when it can be used, on its own, it fails to provide a completely adequate account of trust. This is perhaps not surprising. Each theory of trust necessarily abstracts from the real world and focusses on some more-or-less tractable issue. Problems only really arise when we believe these partial theories to be complete explanations.

2.5.2. The Cognitive Theory of Trust

The cognitive account of trust states that trust arises out of the knowledge and beliefs possessed by an individual[8]. A particular cognitive theory of trust therefore has to say what knowledge and beliefs contribute to trust. Generally theories of interpersonal trust are founded on the view that trust judgments are made on both cognitive and affective grounds[65, 69]. In this light we can see Misztal's perspective on trust is that trust is both cognitive and affective (Passion) with the cognitive aspects being both conscious (Policy) and unconscious (Habitus, at least in part). Our concern here is with a better understanding of the cognitive aspects of trust.

One particular cognitive approach is Potter's *virtue theory of trustworthiness* [89] in which trustworthiness is a virtue that can be recognised by others. To trust *"involves an expectation or belief that the trusted person has good intentions with regard to the care of something we value and the ability to carry through with what is expected of him or her"*. People are trusted because they are trustworthy and we judge trustworthiness by the person demonstrating the virtues, such as *truth telling* or *honesty*, that must accompany trust in their interaction in society and with intimate others. Potter discusses the conditions by which someone can become trustworthy i.e. develop trustworthiness as a virtue. From our perspective she describes a set of beliefs and knowledge of actions that would lead one to believe another is trustworthy.

The simplest beliefs we might require to judge an individual trustworthy are:

- That they are well intentioned in the matter at hand. That is, if they say something, they believe it is true and, in particular, when they promise something or commit to something, they intend to do it.
- That they are competent in the matter at hand. If they say something, it is indeed true or, if they commit to something, they have the skills etc. that are required to do it.

We might also require the additional general beliefs:

- That they do not hide information from us that has a bearing on the matter at hand. In particular that they do not hide information that, had we known it, would have caused us to act differently in the matter at hand.
- That they actually know the necessary information relevant to the situation and their ability to act in it.

[8] And here, inevitably, we run into the deep problem of "inductive generalization" i.e. the generalization from a small number of examples to a general truth. Inductive generalization only has meaning against a background theory that underwrites the generalization by providing some sort of "continuity" condition. When dealing with people, the continuity condition is often the understanding of the underlying motivation for the behaviour. This can go awry when the individual intends to deceive, as in the case of a *sleeper* (i.e. an individual who invests in appearing trustworthy, but has a covert agenda that involves taking advantage of the trust they gain at a later date). Our background theory may rule out such possibilities e.g. on the grounds that the effort involved in creating the subterfuge is disproportionally high compared to any possible gain that may result. However, such an assumption must always be examined carefully since different parties may evaluate the investments required, and gains resulting, differently.

But on top of these there are many specific, situational, beliefs we might require e.g. They are of previous good character. They do not associate with "questionable" characters. They haven't spent extended periods in countries with regimes of which we disapprove without sufficient explanation. The time since their last security vetting has not exceeded the maximum allowed period. And so on. A given situation may require arbitrary specific beliefs about, or knowledge of, an individual before they are regarded as trustworthy. In particular in security applications of the concept of trust these additional aspects can be seen to constitute a major part of what is called the security policy.

Generally beliefs about trusts are beliefs that apply "in normal situations". In times of crisis e.g. in times of war, in economic downturns, when firms go bankrupt, etc. the beliefs on which trust is based are often undermined. Therefore theories of trust often add a specific belief in situational normality which explicitly sanctions the use of the trust supported beliefs[74]. If situational normality does not hold then the beliefs supporting trust should be individually examined to see if they hold in the current circumstances[9].

2.5.3. Trusting Stance

As well as attitudes of trust towards specific individuals, people also have attitudes towards groups or even humanity in general[see again 74]. In some circumstances, for example when we lack specific information, we adopt, either consciously or unconsciously, an attitude of trust or distrust towards others. This is our stance. It may be that we have learnt that the best starting place for a relationship is to start with trust and move to distrust when it is not returned. It may be that we believe that it is better to start with distrust and give trust when it is earned. Our actual position will be the result of Habitus, Passion and Policy and it constitutes a default in our interactions with others.

2.6. Trust and Temporary Systems

In the 1990's business became concerned with the notion of "virtual organisation", teams that come together temporarily for some specific purpose and then dissolve back into the sea of resources to be used again to form new virtual organisations for new purposes. Social scientists classify such dynamic formation and dissolution as *temporary systems* in which individuals may have no direct experience of one another before coming together and have no expectation of ever coming together again.

In "Swift Trust and Temporary Groups", Meyerson, Weick and Kramer discuss the problem of *temporary systems*:

> *"Temporary systems exhibit behavior that presupposes trust, yet traditional sources of trust - familiarity, shared experience, reciprocal disclosure, threats*

[9] It should be clear that normal means *normal for a given context* i.e. what is normal for *this* situation may not be normal for *that* situation. The idea of what is normal therefore changes as background information changes and so both the nature and the extent of trust may change as the result of context changes.

> *and deterrents, fulfilled promises, and demonstrations of non-exploitation of vulnerability - are not obvious in such systems. In this respect, temporary systems act as if trust were present, but their histories seem to preclude its development."*

This poses the problem of how trust forms in such temporary systems given that the usual pre-requisites for "classical trust" seem to be missing. The suggested solution was that a different form of trust, dubbed *swift trust*, came into play where "role and category" information became the main determinants in trust formation. Here role refers to the roles that the members play within the group. Individuals are trusted for the role they play provided that they do not display role violations i.e. inappropriate behaviour for the role. Category refers to the social categorisation performed by group members on other group members and is presumed to be dominated by *"institutional categories that are made salient by the context in which the systems form"*. Members trust one another according to the sharing of perceived categories that the members inhabit. Categories induce a stereotyping of individuals and failure of stereotype expectations may lead to revision of one's trust assessment.

Changes in business practices and the post 90's expansion of the internet have meant that business interest in temporary systems has expanded well beyond the simple virtual organisations of the 90's. Business to business and business to customer interaction are increasingly becoming instances of temporary systems. At the same time as these changes have been taking place in the business environment, corresponding changes have been taking place in governmental, educational and social environments. Consequently, throughout society, increasing numbers of individuals find themselves interacting as parts of temporary systems. Although many of these interactions are mediated by the Internet, many are not. Interactions with institutions, such as banks, governmental bodies and medical practitioners currently predominantly remain face-to-face. But still they cease to be based on long term one-to-one relationships between individuals and instead are performed through an ever changing cast of role holders. In the eyes of many this process, has resulted in institutions becoming "faceless", "bureaucratic" and "soulless" as the "thick relationships" that underwrite "classical trust" become eroded. How can I trust this pensions adviser that I have only just met? How can I trust this doctor that I have never seen before? Progressively each of us is forced either to adopt some new basis for trust or to forgo the advantages offered by co-operation and role differentiation in society. Swift Trust potentially becomes a major bearer of uncertainty reduction in our day to day interactions.

We may see Swift Trust is an example of Luhmann's Generalised Communication Media or Giddens' Abstract Systems. The new abstract system engenders Trust by Role and Category and this is underwritten by systems trust in the Role and Category assignment system. That is, we trust the role occupier because we believe that Role assignment has been carried out by some trustworthy system (which is continually validated by seeking role violations). Similarly Categories are trusted because category assignment is performed on the basis of category tags so that membership of shared categories becomes a basis for trust.

2.7. Trust and Forgiveness

These accounts are somewhat incomplete as they stand, for they fail to account for the role of trust in forgiveness. In many situations there is scope for discretion on the part of the trusted in the fulfilment of obligations. As Baier[5] puts it: *"The more extensive the discretionary powers of the trusted, the less clear-cut will be the answer to the question of when trust is disappointed. The truster, who always needs good judgment to know whom to trust and how much discretion to give, will also have some scope for discretion in judging what should count as failing to meet trust, either through incompetence, negligence, or ill will. In any case of a questionable exercise of discretion there will be room both for forgiveness of unfortunate outcomes and for tact in treatment of the question of whether there is anything to forgive."*

Moreover we cannot assume perfect execution of tasks, there will be disappointment of trust, not because of poor use of discretion on the part of the trustee but because of failures in understanding and failures in performance. Such failures introduce noise into the trust monitoring and enforcement systems (whatever they are). If we regard others as untrustworthy as soon as they fail to meet our expectations then we may expect the set of individuals that we regard as trustworthy to become very small very quickly. Appropriate forgiveness and toleration of error provide a noise immunity mechanism for trustworthiness. Camp et al. [24] observe that there exists a particular asymmetry between forgiveness of failures of trust in intent and trust in competence *"people are ready to forgive harms they may have suffered due to incompetence far more quickly and readily than harms they perceive to have been caused by the bad intentions of others."* (Although we must also observe that there is a limit to the toleration of incompetence. At some point we simply have to regard incompetence as sufficient grounds for no longer trusting in the ability to deliver the desired outcome.)

This need for noise toleration seems to go beyond trust and apply to other aspects of cooperative behaviour. Although a general discussion of this would take us too far afield at this point, the role of forgiveness has been examined in general game theoretic terms in the analysis of *Iterated Prisoner's Dilemma*[4] and *Noisy Iterated Prisoner's Dilemma*[79]. In Iterated Prisoner's Dilemma, forgiveness allows co-operation to be re-established by a TIT-FOR-TAT player after betrayal and 'punishment' by a reciprocal betrayal. In Noisy Prisoner's Dilemma, forgiveness allows GENEROUS-TIT-FOR-TAT to handle acceptable error levels in other players' execution without excessive losses of co-operation.

Trust must also explain the commonplace observation that we are more likely to forgive those we trust than those we distrust. That is, the failure to meet a commitment by someone we trust is often regarded as excusable whereas the same kind of failure by someone we do not actively trust or someone we actively distrust will be seen as (further) evidence of their untrustworthiness. Here we might turn to the theory of social capital popularized in recent years by Robert Putnam's article[92] and subsequent book Bowling Alone[91]. Social capital is defined in analogy to financial capital and is regarded as an investment in social relations that can be an enabler, allowing an individual (or group) to achieve things that they could not otherwise achieve. In this sense trustworthiness can be regarded as social capital that an individual can invest and the possession of social capital can allow the individual to do things and act in particular ways that otherwise would be unacceptable. Although Dasgupta's paper[29], quoted by Gambetta above, does not use the term social capital it does discuss a reputation for trustworthiness as a commodity that may be invested in by an individual. The investment in reputation pays off by creating conditions which allow him to engage in trust based interactions more often. A reputation for trustworthiness also provides a basis for forgiveness. If we recognise that an individual has been trustworthy in the past, even when it has not been to his immediate advantage, then we are more willing to believe that a failure is unintentional and forgivable.

2.8. Summary of Trust

McKnight and Chervany have carried out extensive research on how the term trust is used in trust research literature across disciplines. They have developed frameworks to categorize the different concepts presented, breaking the conceptions down into a common set of component concepts used across the community[71, 72, 73]. Here one of their frameworks[72] is used to summarize the concepts covered in this review.

Trusting Belief

Competence: means having the ability or power to do for one what one needs done

Benevolence: means caring and being motivated to act in one's interest rather than acting opportunistically

Integrity: means making good faith agreements, telling the truth, and fulfilling promises

Predictability: means trustee actions (good or bad) that are consistent enough to be forecasted in a given situation

Dispositional Trust

Faith in Humanity: refers to underlying assumptions about people

Trusting Stance: means that, regardless of what one assumes about other people generally, one assumes that one will achieve better outcomes by dealing with people as though they are well-meaning and reliable.

Institutional Trust[10]

Structural Assurance: means one securely believes that protective structures, guarantees, contracts, regulations, promises, legal recourse, processes, or procedures, are in place that are conducive to situational success

Situational Normality: means one securely believes that the situation in a risky venture is normal or favorable or conducive to situational success.

[10] On other occasions they have used the term *systems trust* for this concept

Trusting Intentions

Willingness to Depend: means one is volitionally prepared to make oneself vulnerable to the other person by relying on them, with a feeling of relative security

Subjective Probability of Depending: means the extent to which one forecasts or predicts that one will depend on the other person, with a feeling of relative security

2.9. Promises To Keep: A Theory of Trust

Later in this dissertation, formal reasoning systems are introduced for *Knowledge on Trust* and *Social Trust*. These formal systems need to be grounded in some particular view of trust which they have some claim to represent.

Here I will offer a simple theory of trust that incorporates a number of the themes surveyed above. This theory of trust is in the realm of policy, of a conscious decision to treat others in a particular manner.

We will define trust as *the rational belief by an individual, the* truster*, that another individual, the* trustee*, will keep a promise where the truster cares about what results from the promise and the truster has chosen to rely on the trustee keeping the promise*.

When in later chapters we may pause and ask: "What kind of thing is this *formal* relation called 'trust' meant to represent?", then the answer is: "the notion of trust described here".

We will define trust as *the rational belief by an individual, the* truster*, that another individual, the* trustee*, will keep a promise where the truster cares about what results from the promise and the truster has chosen to rely on the trustee keeping the promise*.

To understand this definition we must look closely at its parts:

Rational Belief: Belief is not *certain knowledge* but rather arises from incomplete and inconsistent information. Rational Belief is such *uncertain knowledge* supported by *uncertain facts* combined in some manner to give the most plausible, most coherent description of reality that can be assembled from the available information.

For example we may base our assessment of whether or not an individual will keep a promise on a number of factors, such as, the individual's *past behaviour* in similar situations, correlation of the promised behaviour with the general character of the individual (*character*), the degree to which the individual may gain or lose by keeping their promise (*circumstance*) and the *relationship* that exists between us. In general such factors provide neither complete information, nor consistent information, with which to make a decision. So for example we may judge that it is often the case that people break promises when circumstances change in such a manner that they may pay a high penalty if they keep their promise. Yet we might also believe that a person of good character may be expected to keep their promises even in such changed circumstances. Similarly we may hold that the power of a relationship between ourselves and the promiser is such that even if the promiser has a history of breaking promises with other people, we may still expect them to keep promises to *us* .

Promise: A promise is a freely given commitment, made in good faith, between a *promiser* and a *promisee*[11]. Once made a promise can be kept, broken or retracted, but it can only be retracted with the consent of the promisee, i.e. the promisee says "I will not hold you to the promise". Autonomously the promiser can only keep or break the promise. If it is kept only when convenient to the promiser, and broken when it is inconvenient for the promiser to keep it, then it is not a promise.

A promise is *freely given* if it is obtained without *coercion*. It is made in *good faith* if no *deception* has been practiced in order to obtain the promise. A promise obtained by coercion or deception is not a promise at all, i.e. it may be freely retracted when the coercion is removed or the deception is unmasked.

In general, a promise is a commitment to the truth of some proposition. It is often a commitment to a future truth i.e. it will be the case that p is true, and it is often a commitment to an assertion of agency i.e *I will ensure that p*. So *I will do something* is a commitment to eventually *it will be true that I did the thing* i.e. it is commitment to a future truth of a proposition about agency.

Generally, a promise implicitly contains that the promiser both *intends* to keep the promise and that the promiser has the *capabilities* required to keep the promise, although these may be broken out into separate promises. In general we assume that a promiser is always fully aware of his intentions but may be unaware of the limits of their capabilities. And as such we may be willing to forgive a broken promise when it is broken for reason of failure of capability but we are unwilling to forgive a promise broken for reason of failure of intention.

A promise may be implicit as well as explicit. If you tell me something then implicitly there are promises associated with the act of telling. You promise that you do not lie. This does not mean that what you say is true but rather it requires that you are stating something you believe. In this sense you are being *honest*. Moreover, you are also promising you are not lying by omission i.e. you are not withholding beliefs which are relevant to the matter at hand and would lead me to draw different conclusions should you disclose them. We might call this being *disclosing* or *candid*. We may call these the promises of *intention* associated with telling. There are also implicit promises of *performance*. You implicitly promise to be *competent* in a particular domain of knowledge, meaning that if you believe something in that domain then it is true, and to be *knowledgeable* in a particular domain of knowledge, meaning that if something is true in that domain then you will believe it. The implicit performance promises may be waived according to circumstance. For example if I know you are not an expert in the subject

[11] We can extend this to a group of promisers and a group of promisees so that the group of promisers are individually and collectively responsible for the promise to each and every member of the group of promisees.

matter at hand I may waive either or both of the implicit performance promises (or you may caveat the telling by some indication that waives either or both of the performance promises).

Similar principles hold for implicit promises to act in a particular way. When someone offers to act in some way for us we might assume implicitly that they have competencies i.e. the abilities, resources, etc. required to carry out the action, or the knowledge, competence etc. required to guarantee the truth of an assertion.

Moreover, there is no implication of agency on the part of the promiser in the promise, i.e. the promiser is not necessarily the agent that causes the promise to be true (or become true). Consider, *I promise you the sun will rise tomorrow!* Clearly I will not make this happen but the promise has meaning. I am not promising to make the sun rise. I am promising that, in some way, I know the sun will rise tomorrow, even if you doubt it will happen. And if the sun doesn't rise you will feel aggrieved. You will feel I have broken my promise, or at the very least, that I should not have made the promise in the first place.

Finally we have said that a promise is a commitment. What makes a commitment different from, say, a simple observation, is that commitment implies a cost. If I am committed to something I am willing to pay the costs entailed in it being so. There is a big difference between thinking *democracy is a good thing* and being *committed to democracy*. The first is an observation I might make. If circumstances should be about to change to mean the democracy I was living in was about to become a dictatorship then you could not conclude that I would be willing to do anything to oppose it. The second statement however means that I am willing to pay at least some sort of costs to ensure that I go on living in a democracy. If I promise that *the sun will rise tomorrow* I am making a commitment even if the cost is my reputation or our relationship. Here we come close to the contractual meaning of a commitment. If I am contractually committed to something then I am willing, or at least undertake, to pay a forfeit if I do not meet that commitment.

If there is no promise there is no trust. There is simply a belief that in certain circumstances someone will do something and in other circumstances they will not. There may indeed be reliability in this prediction and statistical or other reasons backing up the belief. But this only means that there is predictability rather than trust.

Cares: If the potential truster simply doesn't care about what results from the promise then there is no need for trust.

Chooses: The potential truster may care about the outcome and believe the trustee will keep the promise. But the potential truster may choose not to trust the potential trustee and take other steps to manage the outcome, such as hedging or insurance. Here we acknowledge that the truster is also autonomous and that trust, or indeed trusting *this particular individual*, is not the only option open to him. The potential truster makes choices between

> alternatives not only on the basis of outcomes but also on other grounds such as the *morality of the situation* and the basis of *how outcomes are achieved*. For example, I may choose not to trust someone and may take alternative steps to guarantee an outcome because I do not like the methods that the potential trustee is likely to employ. I may well believe that they may be trustworthy and that I may care about the outcome, but I would rather the outcome was not achieved by *their* agency. In choosing not to trust you I decline your offer and reject any reciprocal obligation to you in the matter at hand.
>
> If the potential truster cares about the outcome and has no choice but to rely on the trustee then there is no issue of trust i.e trust involves the voluntary surrender of control to another individual.

An individual regards another as trustworthy if they believe that the other will keep their promises. The central question of trust and trustworthiness thus becomes: What is sufficient reason to believe another will keep their promises?

2.9.1. Classical Trust

For classical trust this belief is built from experience of the past behaviour of the individual and an understanding of their motivations. We might give an account of this as follows.

The judgment of trustworthiness is based upon specific knowledge of, and beliefs about, an individual. Part of this is knowledge of past behaviour and beliefs about the motivations for that behaviour. A sufficient reason to believe an individual is trustworthy is the belief that the individual behaves in a particular manner to maintain a chosen property, together with the knowledge of, or belief about, a particular mechanism that will lead to the property being damaged if the individual fails to live up to promises they have made.

We will refer to this combination of property and mechanism as the motivation of the individual to live up to their promises. Additionally systems trust may be required to support the mechanism, when, for example, the mechanism is an institutional framework e.g. a professional association whose membership constitutes a certified body of practitioners. However the mechanism may incorporate a degree of forgiveness. Not all failures to live up to promises will result in immediate punishment. Generally the degree of toleration in the mechanism will also be based on knowledge and beliefs about the trustee, and this knowledge and belief is updated as a result of failure so that different consequences may occur in the future. The trustor will also update their beliefs as a result of fulfilment, and failure of fulfilment of obligations by the trustee. But the trustor will update beliefs about the motivations of the trustee i.e. beliefs about the property being maintained by the trustee and beliefs about the mechanisms that damage this property. These revised beliefs may lead the trustor to make different trust decisions in the future, both of the particular trustee that caused the update and potential trustees in general.

Trusting stance enters the model as the choice made for the default motivation model when we lack specific information. Figure 2.2 summarises this diagrammatically.

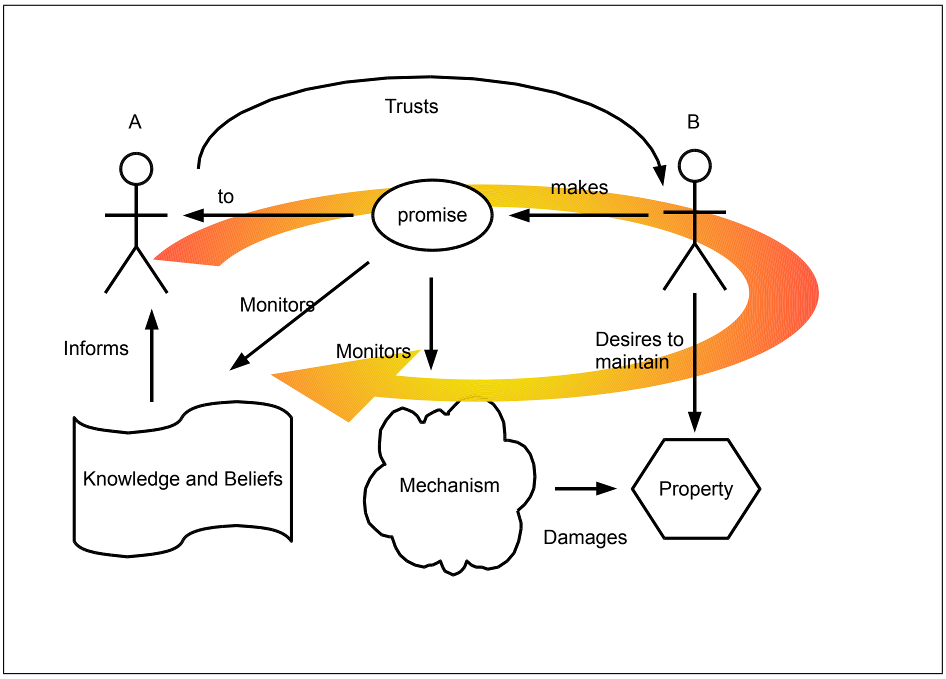

Figure 2.2.: Classical Trust Model

Subjective probability re-enters the picture as an assessment of the trustor's confidence in the supposed motivation of the trustee. That is, it re-enters the picture as the specific process of evaluating the evidence associated with a particular explanation of why the trustee is trustworthy.

2.9.2. Swift Trust

But what of the cases where we do not have direct experience of the individual? What of trust in temporary systems? As indicated in the earlier discussion of temporary systems, swift trust offers an answer and we may explain trust based on role and category in terms of a similar model. If a person is in a role, R, then promises made in that role are regulated by a mechanism, M, that penalises role holders that fail to fulfil those particular promises. One particular property that the role holder might wish to maintain is the ability to inhabit the given role in the future and the enforcement mechanism that might apply is role exclusion in that individuals that fail to fulfil the promises associated with a role become excluded from holding that role in the future. Similarly, if a person really is in a category, C, then there is a mechanism that will penalise them for violations of the promises associated with C. If we have systems trust in the role and category assignment mechanisms then we can extend trust to individuals in particular roles and categories because of the motivations associated with role and category. The difference between Swift Trust and Classical Trust can be seen by comparing the Classical Trust diagram, figure 2.2, and the Swift Trust Diagram figure 2.3.

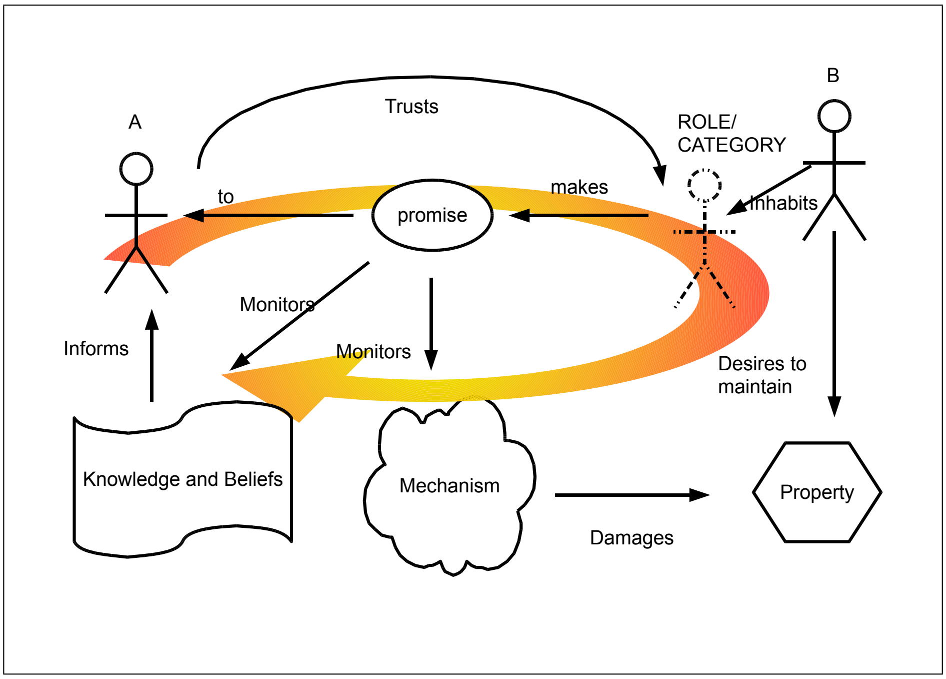

Figure 2.3.: Swift Trust Model

2.9.3. Social Embeddedness and Indirect Trust

However the above are not the only ways we obtain belief that others will keep their promises. We are also *told* things by people we trust. We are told that people are trustworthy, or untrustworthy. We are told about their circumstances, we are told about their past behaviour and we are told about their virtues (and failings). Each and every one of us is embedded in a network of established and long lasting relationships. These networks allow us to bootstrap initial trust in people we do not directly know using referrals from our network. Our social embeddedness offers an alternative and complementary approach to swift trust, giving us another way to establish trust in temporary systems. I have initial trust in you because of what I have been told by intermediaries, and this initial trust is underwritten by my trust in the intermediaries.

This leads to the consideration of two different problems which depend on what kind of information I am given by intermediaries.

The first problem arises where my social network gives me information about the circumstances surrounding events and individuals. The information I receive may be both incomplete and inconsistent, and the sources I receive of information may not be all equally trustworthy. In this case the problem is to synthesise the most plausible picture of the situation from the information I have received given the relative trustworthiness of the sources. This will be called the problem of *Knowledge on Trust*. This problem is the subject of Chapter 4.

In the second problem I am given *referrals* which say whether an individual is trustworthy or untrustworthy. But the problem is reflexive in that referees may indicate not just on their beliefs about the individual of concern but also about other referees. This will be called the problem of *Social Trust*.

It should be emphasised that this is not a theory of transitive trust. Rather it is a theory of piecing together different trust relationships to form an overall picture of trust by using the promises made by others. I trust A as a referee. A recommends B as a referee as to whether C is trustworthy for some task. I am willing to accept the chain because I trust A's judgement about B and B's judgement about C and both have made appropriate promises. If I believed only A trusts B and B trusts C in the appropriate ways this would not be enough for me to trust C. The recommendations themselves are an important part of establishing my (initial) trust in C because they carry the promises implicitly within them.

A central part of the problem is resolving the cancellations that arise when one distrusts: i.e. if a referee is distrusted then the referee's distrust of another is "'cancelled out". This does not mean that the other is trusted but that the referee's opinion is discounted. The resolution of an opinion synthesised from networks of referees that trust and distrust one another is the problem is taken up in Chapter 5.

3. The Mathematical Framework, including a Primer on Binary Relations and Galois Connections

3.1. Introduction

The computer scientist approaches problems with a particular outlook and a particular toolkit. Part of that outlook, shared with other sciences, such as physics, is to approach problems by creating formal models of them. Formal models allow us to abstract from the detail of situations and processes and explore how elements of a problem may interact. Another, and very important, part of the outlook is to seek out the computational content of a concept. Modelling should not only result in a description of some information process but also a hypothesis about how the process may actually be performed effectively.

So far this dissertation has explored the concept of trust in the same manner as might be pursued in a philosophical or sociological context. This chapter marks a transition from this approach into one in which we start to use some of the tools from computer science to explore the concept of trust, or at least aspects of it, more rigorously. The following chapters will adopt a model building approach to the concepts of Knowledge-on-Trust and Social-Trust introduced earlier. The style adopted is to mix informal discussion of ideas with their formalisation in simple set theory and the development of computational methods to address the respective concepts.

The use of mathematics becomes more pronounced as the dissertation progresses, so that in Chapter 4 the mathematical notation is only used to specify the requirements of a program to compute coherent theories whereas in the following chapters the formal mathematical framework is used both to specify notions such as argumentation and trust system and analyse these notions through proofs of properties.

In subsequent chapters we regard trust as being associated with logical problems, not in the sense of deductive inference, but rather in the sense of finding solutions under logical constraints. To illustrate, consider a problem from Raymond Smullyan's "What is the Name of this Book?" [106].

> *In Shakespeare's Merchant of Venice Portia had three caskets, gold, silver, and lead, inside one of which was Portia's portrait. The suitor was to choose one of the caskets, and if he was lucky enough (or wise enough) to choose the one with the portrait, then he could claim Portia as his bride. On the lid of each casket was an inscription to help the suitor choose wisely.*
>
> *Now, suppose Portia wished to choose her husband not on the basis of virtue,*

> *but simply on the basis of intelligence. She had the following inscriptions put on the caskets.*
>
> ***Gold:*** *The portrait is in this casket.*
>
> ***Silver:*** *The portrait is not in this casket.*
>
> ***Lead:*** *The portrait is not in the Gold casket.*
>
> *Portia explained to the suitor that of the three statements, at most one was true.*
>
> *Which casket should the suitor choose?*[1]

This problem is solved, not by deductive inference, but by determining which assignments of true and false to statements causes all the conditions to be met. The problem may be translated into propositional logic in several ways. A fairly literal translation is given by naming each alternative proposition as to which casket holds the portrait:

$$PG \text{ is The portrait is in the Gold casket.} \tag{3.1}$$

$$PS \text{ is The portrait is in the Silver casket.} \tag{3.2}$$

$$PL \text{ is The portrait is in the Lead casket.} \tag{3.3}$$

Stating that only one casket can be holding the portrait (i.e. there is only one portrait):

$$(PG \wedge \neg\, PS \wedge \neg\, PL) \vee (\neg\, PG \wedge PS \wedge \neg\, PL) \vee (\neg\, PG \wedge \neg\, PS \wedge PL). \tag{3.4}$$

Naming each of the propositions written on the casket and define what they say about the portrait location:

$$G \equiv PG. \tag{3.5}$$

$$S \equiv \neg\, PS. \tag{3.6}$$

$$L \equiv \neg\, PG \tag{3.7}$$

And stating that only one of these statements is true:

$$(G \wedge \neg\, S \wedge \neg\, L) \vee (\neg\, G \wedge S \wedge \neg\, L) \vee (\neg\, G \wedge \neg\, S \wedge L). \tag{3.8}$$

The problem thus becomes the problem of assigning the values true and false to each of the propositions PG, PS, PL, G, S and L in such a manner that the various formulae are all true. Such an assignment is called a (propositional) model of the collection of formula. The mechanical approach to this is to use a model checker such as MACE4 that will systematically search through the assignments to determine which satisfy all the formulas. In this case, for example, MACE4 results in exactly one possible model being found:

[1] Silver.

$$\begin{array}{lll} G = \textbf{false} & G = \textbf{false} & L = \textbf{true} \\ PG = \textbf{false} & PS = \textbf{true} & PL = \textbf{false} \end{array}$$

Below the basic framework of propositional model finding is given. Chapter 4 uses an extended version of propositional model checking to address the problem of Knowledge-on-Trust and Chapter 5 shows how questions of Social-Trust may be translated into the propositional model checking framework. In pursuing this latter goal, binary relations, operators derived from binary relations and Galois Connections are used.

Binary relations are familiar to many computer scientists as a representation of non-deterministic programs with operations on binary relations being ways of combining non-deterministic programs or program fragments[2]. Similarly operators derived from binary relations are familiar as predicate transformers such as weakest (liberal) precondition and strongest postcondition[3]. The notion of a Galois Connection arises in the laws that connect predicate transformers. So, for example, given a program, R, the weakest liberal precondition of R for result Q, written $\mathrm{wlp}(R, Q)$ is the largest set, P, such that if one runs R with input in P, and R terminates, then the result is in Q. The predicate transformer is the operator $\mathrm{wlp}(R)$ obtained by abstracting over Q. The strongest postcondition, written $\mathrm{sp}(R, P)$, is the dual such that Q is the smallest set such that starting in P and terminating one is guaranteed to be in Q. The predicate transformer, of course, being $\mathrm{sp}(R)$. In this case, the Galois Connection, written in terms of the operators, is the law that connects the weakest liberal precondition and strongest post condition $\mathrm{sp}(R)P \subseteq Q \equiv P \subseteq \mathrm{wlp}(R)Q$. The framework of binary relations, operators and Galois Connections will be used in the development of argumentation systems and their use as a tool for modelling Social-Trust in

[2] Exactly when programs were seen as relations, and by whom, seems hard to pin down. Certainly by the Royal Society workshop in 1984 Tony Hoare could clearly portray programs as predicates [54], a view he attributes initially to Rick Hehner during Hehner's extended visit to the PRG. And from predicates it is but a short step to relations. But this is far from the whole story. John Sanderson's monograph "A relational Theory of Computing" [98] was published in 1980 and presents a fully relational semantics for a small language and Sanderson himself was aware of Andrezej Blikle's work on using relations in studying program analysis techniques [15]. Moreover, Blikle subsequently went on to develop MetaSoft, a set based approach to denotational semantics [16]. Likewise one might say that the relational approach is implicit in much of Jean-Raymond Abrial's work on the set theoretic approach to specification that eventually, after much mutation, became Z. We should also note that at approximately the same time the world of domain theoretic semantics was also active in the area of the relational semantics of programming language with, for example, Mike Smyth publishing in 1983 a power domain approach to continuous multifunctions and predicate transformers [107]. Perhaps the reality is that there was a gradual movement of ideas from their use as a tool for specifying programming languages to a tool for specifying programs, and the gradual realisation that relations and predicates were but two sides of the same coin.

[3] Predicate Transformers and weakest Preconditions were introduced by Edsger W. Dijkstra and the canonical references are Dijkstra's "Guarded commands, nondeterminacy and formal derivation of programs" [31] and book "A Discipline of Programming" [32], with the more up-to-date, and more formal, reference being Dijkstra's and Carl S. Scholten's "Predicate Calculus and Program Semantics" [33]. The exact details of how one represents non-termination of a program are an issue in deciding exactly what is the relational representation of a program. Here we are intentionally inexact but if the matter is of concern to the reader we may limit the discussion to always terminating programs

subsequent chapters.

This chapter next provides a brief review of the main mathematical tools that are used in subsequent chapters.

3.2. Informal Set Theory

The mathematical discussion in this dissertation is conducted in an informal set theory with typing[4]. The set theory can be regarded as a fragment of the Z specification language (see, for example, Mike Spivey's "Z Notation - a reference manual (2. ed.)" [110]). We take as base types a potentially infinite collection of *sorts*, named A, B, C etc. and take type constructors power, $\mathbb{P}\,\tau$, and (binary)product $\tau \times \sigma$ for types τ and σ. Each sort corresponds to a non-empty set of the same name. Power set and the product of two sets are likewise denoted $\mathbb{P}\,T$ and $T \times S$ for sets T and S. Given a type declaration, $s : \tau$ we assume the assertion $s \in T$ where T is the obviously corresponding set to τ. Set comprehension is written $\{x : S \mid \mathit{predicate}\}$ or $\{\mathit{term} \mid \mathit{predicate}\}$ in the conventional mathematical style. When we wish to emphasise type membership rather than set membership in a declaration we will write $x : S$, rather than $x \in S$, to emphasise that S is a type, often, however, where the type may be determined implicitly from the context, we will write neither. Operators such as union ($\cup$), intersection ($\cap$), set difference ($\setminus$) and complement ($\overline{\_}$) are taken as defined in all appropriate types (complement being defined with respect to the type i.e. for $X : \mathbb{P}\,\tau$, complement is defined by $\overline{X} \mathrel{\widehat{=}} \{x : \tau \mid x \notin X\}$). In future when operations are introduced they will be available in all appropriate types without explicit comment.

3.3. Propositional Logic and Propositional Models

We now formally discuss propositional model finding. First we discuss propositional logic as a formal language giving the meaning of collections of propositions in term of sets of assignments of true or false to basic propositions, which we call propositional variables (or simple variables). Second we illustrate the process of model finding by defining a non-deterministic program (relation) for finding a model[5].

The language of propositional logic is that of the connectives "*and*", "*or*", "*not*" and their derivatives. It is concerned with tying together atomic propositions, which are unanalysed statements, which are either true or false. These atomic unanalysed statements will be modelled by some suitably large collection of variables. The language of propositional logic then lets us build propositional sentences out of these variables according to the following grammar with the connectives taking their usual meaning:

[4] Or, if one prefers to so call it, a simple type theory.

[5] The material in this section is loosely based on Raymond Smullyan's book "First Order Logic" [105] with the model finding procedure being in essence a procedure for non-deterministically finding an open branch of a semantic tableau. Other works presenting a similar, tableaux based, view of logic include the introductory books by Hodges [55], Priest [90] and Jeffrey [58].

$$
\begin{array}{lll}
\mathit{Proposition} & ::= & \mathit{Proposition} \wedge \mathit{Proposition} \mid \\
 & & \mathit{Proposition} \vee \mathit{Proposition} \mid \\
 & & \mathit{Proposition} \Rightarrow \mathit{Proposition} \mid \\
 & & \mathit{Proposition} \equiv \mathit{Proposition} \mid \\
 & & \neg\ \mathit{Proposition} \mid \\
 & & (\mathit{Proposition}) \mid \\
 & & \mathbf{true} \mid \\
 & & \mathbf{false} \mid \\
 & & \mathit{Variable}
\end{array}
$$

Given a set, S, of the propositional variables, an assignment for S is a function from S to $\{0,1\}$. An assignment represents which variables are taken to be true (assigned 1) and which are taken to be false (assigned 0). The semantics of propositional logic extends the notion of truth of variables to the notion of truth of sentences. Given a propositional sentence, P, an assignment for P is an assignment, $m : S \to \{0,1\}$, such that S includes all the variables mentioned in P. An assignment a for P, is a model of P, written $a \models P$ if, and only if, it obeys the following recursive condition:

$$
\begin{array}{lll}
m \models p \wedge q & \textbf{iff} & m \models p \ \textbf{and}\ m \models q \\
m \models p \vee q & \textbf{iff} & m \models p \ \textbf{or}\ m \models q \\
m \models p \Rightarrow q & \textbf{iff} & m \not\models p \ \textbf{or}\ \ m \models q \\
m \models p \equiv q & \textbf{iff} & m \models p \ \textbf{iff}\ \ m \models q \\
m \models \neg\, p & \textbf{iff} & m \not\models p \\
m \models \mathbf{true} & & \\
m \not\models \mathbf{false} & & \\
m \models v & \textbf{iff} & m(v) = 1 \text{ for } v \text{ a variable}
\end{array}
$$

where $m \not\models x$ means **not** $m \models x$.

Given a set of propositional sentences an assignment is a model for the set if it is a model of every sentence in the set. Given a set A of propositional sentences the set of all models of A is $\mathrm{Mod}(A) \mathrel{\widehat{=}} \{m \mid \forall a \in A. m \models a\}$. The set A is said to be consistent if $\mathrm{Mod}(A)$ is non-empty, otherwise it is said to be inconsistent. The set A is said to be consistent with a propositional sentence p if there is an assignment that is simultaneously a model of both A and p, or equivalently, $\mathrm{Mod}(A \cup \{p\})$ is non-empty. A propositional sentence, p, is a consequence of the set A if there are no assignments that are simultaneously a model of $\neg\, p$ and A, i.e. $\mathrm{Mod}(A \cup \{\neg\, p\})$ is empty[6].

Since a set of propositional sentences may have many models, one can impose extra selection conditions to pick out particular models or sets of models. For example, we may select models because we prefer certain variables to take on the value 1 (or 0) over others. Or we may select models that maximise (or minimise) the number of mentioned variables that take on the value 1. Such additional constraints will occur when propositional models are used to give computational semantics in later chapters.

We now turn to the matter of finding models for a set of sentences. There

[6] That is, it is not possible for everything in A to be true and p not to be true.

are many possible algorithms for doing this. Here, by way of illustration, we give a non-deterministic algorithm for finding a model. All models are found by pursuing all possible choices at choice points rather than a single option.

Let an atomic proposition be a propositional variable or negated propositional variable. An atomic proposition p is said to be positive occurrence of v if p is the propositional variable v and said to be a negative occurrence of v if p is the negation of v. A model set is a set, S, of atomic propositions such that no variable occurs as both a positive and negative occurrence in S. Model sets are a coding of assignments to propositional variables with a variable v being assigned 1 if it occurs positively in the set and assigned 0 if it occurs negatively. Given a non-empty model set, if a variable does not occur in the model set then it may be assigned either 1 or 0.

Given a set of propositional sentences S we define the non-deterministic reduction relation, $\leadsto$, on S by:

$$\begin{array}{lll}
\{\alpha\} \cup \Gamma & \leadsto & \alpha_1 \cup \Gamma \\
\{\beta\} \cup \Gamma & \leadsto & \beta_1 \cup \Gamma \\
\{\beta\} \cup \Gamma & \leadsto & \beta_2 \cup \Gamma \\
\{v\} \cup \{\neg\, v\} \cup \Gamma & \leadsto & \emptyset \text{ for } v \text{ a variable} \\
\{\mathbf{false}\} \cup \Gamma & \leadsto & \emptyset
\end{array}$$

where

α	α_1
$a \wedge b$	$\{a, b\}$
$\neg\,(a \vee b)$	$\{\neg\, a, \neg\, b\}$
$\neg\,(a \Rightarrow b)$	$\{a, \neg\, b\}$

β	β_1	β_2
$a \vee b$	$\{a\}$	$\{b\}$
$\neg\,(a \wedge b)$	$\{\neg\, a\}$	$\{\neg\, b\}$
$(a \Rightarrow b)$	$\{\neg\, a\}$	$\{b\}$
$(a \equiv b)$	$\{a, b\}$	$\{\neg\, a, \neg\, b\}$
$\neg\,(a \equiv b)$	$\{a, \neg\, b\}$	$\{\neg\, a, b\}$

S is irreducible iff $\neg\,\exists\, X . S \leadsto X$.

M is a model set for S iff $S = M$ or there is a finite sequence of reductions $S = X_1 \leadsto X_2 \leadsto \ldots \leadsto X_{n-1} \leadsto X_n = M$ and M is irreducible and non-empty. If this is the case we will write $S \leadsto^\sharp M$.

It is straightforward to show that if $X \leadsto X'$ and $m \models X'$ then $m' \models X$ where m' is an extension of m[7] and so, by induction, that if $X \leadsto^\sharp X'$ and $m \models X'$ then $m' \models X$.

If one views $\leadsto^\sharp$ via predicate transformers one can see that $\mathrm{sp}(\leadsto^\sharp)(\{p\})$ is the set of all models of the proposition p and given a model set m, $\mathrm{wlp}(\leadsto^\sharp)m$ is the set of all propositional sentences that have m as a model set[8].

3.4. Binary Relations

Binary relations from A to B are the subsets of $A \times B$. If R is a binary relation we write $R : A \leftrightarrow B$ as an alternative to $R : \mathbb{P}(A \times B)$ and xRy as an alternative to

[7] Since the reduction step may throw away a variable that occurred in X, the assignment m may not fully determine the variables in X and an arbitrary choice must be made for the assignment to variables so thrown away.

[8] Of course, in this latter process the β steps can introduce an arbitrary number of new variables that may be arbitrarily true or false. And even without this potential source of infinity, even the α steps can introduce semantic duplication of formulas, e.g. p and $p \wedge p$.

$(x, y) \in R$. Here we give a brief overview of binary relations and their properties[9].

Given a binary relation $R : A \leftrightarrow B$, the derived relations $\breve{R}$ and $\overline{R}$ of converse and complement relations are defined by $x\breve{R}y \mathrel{\widehat{=}} yRx$ and $x\overline{R}y \mathrel{\widehat{=}} \neg\, xRy$ where x ranges over A and y over B, and again we emphasise that complementation is taken relative to $A \times B$. Given two binary relations $R : A \leftrightarrow B$ and $S : B \leftrightarrow C$ the operations of intersection and union, inherited from sets, create new binary relations and the relational composition of R and S, denoted $R \fatsemi S$ is defined by $xR \fatsemi Sy \mathrel{\widehat{=}} \exists\, z.xRz \wedge zSy$ (the "generalised functional composition" is defined as $S \circ R \mathrel{\widehat{=}} R \fatsemi S$) where x ranges over A, z over B and y over C. The identity relations $I : A \leftrightarrow A$ is defined by $xIy \mathrel{\widehat{=}} x = y$, where x and y range over A, and the universal, or chaos, relation, $U : A \leftrightarrow B$, is defined by $xUy = \textbf{true}$ where x ranges over A and y over B.

The *natural* domain and *natural* range of $R : A \leftrightarrow B$ are the first and second projections of the relation R defined by $pr_1R = \{x : A \mid \exists\, y : B.xRy\}$ and $pr_2R = \{y : B \mid \exists\, x : A.xRy\}$.

The language of binary relations is very expressive and many properties of relations may be stated directly in terms of the algebra of relations without resort to the underlying set theory. As examples consider the following commonly used properties of binary relations:

property	**definition**	
$R : A \leftrightarrow B$ is a *partial function*	$\breve{R} \fatsemi R \subseteq I_B$	(3.9)
$R : A \leftrightarrow B$ is *total*	$I_A \subseteq R \fatsemi \breve{R}$	(3.10)
$R : A \leftrightarrow B$ is a *surjection*	$I_B \subseteq \breve{R} \fatsemi R$	(3.11)
$R : A \leftrightarrow B$ is an *injection*	$R \fatsemi \breve{R} \subseteq I_A$	(3.12)
$R : A \leftrightarrow A$ is *reflexive*	$I_A \subseteq R$	(3.13)
$R : A \leftrightarrow A$ is *irreflexive*	$R \subseteq \overline{I_A}$	(3.14)
$R : A \leftrightarrow A$ is *symmetric*	$\breve{R} = R$	(3.15)
$R : A \leftrightarrow A$ is *asymmetric*	$R \subseteq \overline{\breve{R}}$	(3.16)
$R : A \leftrightarrow A$ is *antisymmetric*	$R \cap \breve{R} \subseteq I_A$	(3.17)
$R : A \leftrightarrow A$ is *transitive*	$R \fatsemi R \subseteq R$	(3.18)
$R : A \leftrightarrow A$ is *negatively-transitive*	$\overline{R} \fatsemi \overline{R} \subseteq \overline{R}$	(3.19)
$R : A \leftrightarrow A$ is *connected*	$U_A \subseteq R \cup \breve{R}$	(3.20)
$R : A \leftrightarrow A$ is *weakly connected*	$\overline{I_A} \subseteq R \cup \breve{R}$	(3.21)
$R : A \leftrightarrow A$ is an *equivalence*	$\breve{R} \fatsemi R = R$	(3.22)

The operators $\cup, \cap, \bar{}$ form the usual boolean algebra, inherited from sets, over relations. The remaining operators interact with these boolean operators and between themselves in various ways. In later chapters we will use some of these interactions in proving theorems. The following are some of the useful theorems about the algebra of binary relations, including how relational composition distributes over unions and intersections of relations:

[9] The particular properties being largely taken from Ono's "some Properties of Binary Relations" [82] and Riguet's "Relations binaires, fermetures, correspondances de Galois" [93].

assumptions	**theorem**	
$R : A \leftrightarrow B$	$I_A \mathbin{;} R = R = R \mathbin{;} I_B$	(3.23)
$R : A \leftrightarrow A$	$R \subseteq I_A \Rightarrow \breve{R} = R = R \mathbin{;} R$	(3.24)
$\forall k \in K.R_k : A \leftrightarrow B$	$\breve{(\bigcup_k R_k)} = \bigcup_k \breve{R}_k$	(3.25)
$\forall k \in K.R_k : A \leftrightarrow B$	$\breve{(\bigcap_k R_k)} = \bigcap_k \breve{R}_k$	(3.26)
$R, S : A \leftrightarrow B$	$\breve{(R \setminus S)} = \breve{R} \setminus \breve{S}$	(3.27)
$R : A \leftrightarrow B$	$R = \breve{\breve{R}}$	(3.28)
$R : A \leftrightarrow B$	$R = \overline{\overline{R}}$	(3.29)
$R, S : A \leftrightarrow B$	$\breve{(R \mathbin{;} S)} = \breve{S} \mathbin{;} \breve{T}$	(3.30)
$R, S, T : A \leftrightarrow B$	$R \mathbin{;} (S \mathbin{;} T) = (R \mathbin{;} S) \mathbin{;} T$	(3.31)
$R : A \leftrightarrow B, \forall k \in K.S_k : A \leftrightarrow B$	$R \mathbin{;} (\bigcup_k S_k) = \bigcup_k (R \mathbin{;} S_k)$	(3.32)
$\forall k \in K.R_k : A \leftrightarrow B, S : A \leftrightarrow B$	$(\bigcup_k R_k) \mathbin{;} S = \bigcup_k (R_k \mathbin{;} S)$	(3.33)
$R : A \leftrightarrow B, \forall k \in K.S_k : A \leftrightarrow B$	$R \mathbin{;} (\bigcap_k S_k) \subseteq \bigcap_k (R \mathbin{;} S_k)$	(3.34)
$\forall k \in K.R_k : A \leftrightarrow B, S : A \leftrightarrow B$	$(\bigcap_k R_k) \mathbin{;} S \subseteq \bigcap_k (R_k \mathbin{;} S)$	(3.35)
$R, S, T : A \leftrightarrow B$	$(R \setminus S) \mathbin{;} T \supseteq (R \mathbin{;} T) \setminus (S \mathbin{;} T)$	(3.36)
$R, S, T : A \leftrightarrow B$	$R \mathbin{;} (S \setminus T) \supseteq (R \mathbin{;} S) \setminus (R \mathbin{;} T)$	(3.37)
$R : A \leftrightarrow B$	$R \subseteq R \mathbin{;} \breve{R} \mathbin{;} R$	(3.38)
$R : A \leftrightarrow B$	$R = R \mathbin{;} (\breve{R} \mathbin{;} R \cap I_B)$	(3.39)
$R : A \leftrightarrow B$	$R = (R \mathbin{;} \breve{R} \cap I_A) \mathbin{;} R$	(3.40)
$R, S : A \leftrightarrow B$	$R \subseteq S \Rightarrow \breve{R} \subseteq \breve{S}$	(3.41)
$R, S : A \leftrightarrow B, T : B \leftrightarrow C$	$R \subseteq S \Rightarrow R \mathbin{;} T \subseteq S \mathbin{;} T$	(3.42)
$R, S : A \leftrightarrow B, T : C \leftrightarrow A$	$R \subseteq S \Rightarrow T \mathbin{;} R \subseteq T \mathbin{;} S$	(3.43)

If we introduce the notion of the powers, or iterates, of a relation, $R : A \leftrightarrow A$ defined inductively by:

$$R^0 \mathrel{\widehat{=}} I \qquad (3.44)$$
$$R^1 \mathrel{\widehat{=}} R \qquad (3.45)$$
$$R^{n+1} \mathrel{\widehat{=}} R \mathbin{;} R^n \qquad (3.46)$$
$$R^+ \mathrel{\widehat{=}} \bigcup_{i \geq 1} R^i \qquad (3.47)$$
$$R^* \mathrel{\widehat{=}} \bigcup_{i \geq 0} R^i \qquad (3.48)$$

Then we also have:

assumptions	**theorem**	
$R : A \leftrightarrow A$	$R^m \mathbin{;} R^n = R^{m+n}$	(3.49)
$R, S : A \leftrightarrow A$	$R \subseteq S \Rightarrow R^n \subseteq S^n$	(3.50)
$R : A \leftrightarrow A$	$R \mathbin{;} R^* = R^+$	(3.51)
$R : A \leftrightarrow A$	$R^* \mathbin{;} R = R^+$	(3.52)
$R : A \leftrightarrow A$	$R^* \mathbin{;} R^* = R^*$	(3.53)
$R : A \leftrightarrow A$	$I_A \cup R^+ = R^*$	(3.54)
$R : A \leftrightarrow A$	$R^* \mathbin{;} R^n \subseteq R^*$	(3.55)
$R : A \leftrightarrow A$	$R^n \mathbin{;} R^* \subseteq R^*$	(3.56)
$R : A \leftrightarrow A$	R^+ is transitive	(3.57)
$R : A \leftrightarrow A$	R^* is reflexive and transitive	(3.58)

As an example of the notion of iterated relations we see that $\rightsquigarrow^\sharp \subseteq \rightsquigarrow^*$, and that $\rightsquigarrow^\sharp$ is $\rightsquigarrow^*$ restricted to $(\mathbb{P}\,\text{Propositions} \leftrightarrow \mathbb{P}\,\text{Atomic Propositions})$.

3.5. Operators

When dealing with binary relations of the form $R : A \leftrightarrow B$ we will use the term operator to mean a function $f : \mathbb{P}\,A \to \mathbb{P}\,B$. Certain operators derived from binary relations provide a very useful alternative way of dealing with relations by providing what we might call a *functional view* of relations. In this section and the next we will introduce the idea of an operator derived from a relation and the notion of a closure property. In the subsequent section we will formally introduce Galois Connections and the Galois Connections derived from a relation between sets[10]. Given $R : A \leftrightarrow B$ we define *the evaluation of* R *at a point*, denoted $R.x$, as $\{y \mid xRy\}$. Using this evaluation at a point we can readily define two operators associated with R by:

$$R_\Sigma X \mathrel{\widehat{=}} \bigcup_{x \in X} R.x \tag{3.59}$$

$$R_\Pi X \mathrel{\widehat{=}} \bigcap_{x \in X} R.x \tag{3.60}$$

The relation is recoverable from each of the operators by:

$$y \in R_\Sigma\{x\} \equiv xRy \equiv y \in \overline{\overline{R}_\Pi\{x\}} \tag{3.61}$$

The behaviour of these operators on sets is determined by their behaviour on points:

$$R_\Sigma(X \cup Y) = (R_\Sigma X) \cup (R_\Sigma Y) \tag{3.62}$$

$$R_\Pi(X \cup Y) = (R_\Pi X) \cap (R_\Pi Y) \tag{3.63}$$

and:

$$X \subseteq Y \Rightarrow R_\Sigma X \subseteq R_\Sigma Y \tag{3.64}$$

$$X \subseteq Y \Rightarrow R_\Pi X \supseteq R_\Pi Y \tag{3.65}$$

$$\tag{3.66}$$

And since the behaviour of these operators on sets is determined by their behaviour on points we have:

[10] The material for these three sections is extensively based on Ore's paper"Galois Connexions" [83] and book "Theory of Graphs" [84], Everett's "Closure Operators and Galois Theory in Lattices" [38], Erne, Koslowski, Melton and Strecker's "A Primer on Galois Connections" [37], Erne's "Adjunctions and Galois Connections: Origins, History and Development" [36] and Birkhoff's book "Lattice Theory" [14].

$$(\forall\, x : A.R.x = S.x) \Rightarrow (\forall\, X : \mathbb{P}\, A.R_{\Sigma}X = S_{\Sigma}X) \tag{3.67}$$
$$(\forall\, x : A.R.x = S.x) \Rightarrow (\forall\, X : \mathbb{P}\, A.R_{\Pi}X = S_{\Pi}X) \tag{3.68}$$

and

$$(\forall\, X.R_{\Sigma}X = S_{\Sigma}X) \Rightarrow R = S \tag{3.69}$$
$$(\forall\, X.R_{\Pi}X = S_{\Pi}X) \Rightarrow R = S \tag{3.70}$$

R_{Σ} and R_{Π} are readily seen to be dual to one another by:

$$\overline{\overline{R}_{\Sigma}X} = R_{\Pi}X \qquad\qquad \overline{\overline{R}_{\Pi}X} = R_{\Sigma}X \tag{3.71}$$

and we may equivalently express R_{Σ} and R_{Π} as

$$R_{\Sigma}X = \{y \mid \exists\, x.x \in X \wedge xRy\} \tag{3.72}$$
$$R_{\Pi}X = \{y \mid \forall\, x.x \in X \Rightarrow xRy\} \tag{3.73}$$

Some of the useful properties of operators are tabulated below:
Assume $X \subseteq A$, $Y \subseteq A$ and for all $k \in K$, $X_k \subseteq A$,

assumptions	Σ	Π	
$R, S : A \leftrightarrow B,$	$R \subseteq S \Rightarrow R_{\Sigma}X \subseteq S_{\Sigma}X$	$R \subseteq S \Rightarrow R_{\Pi}X \subseteq S_{\Pi}X$	(3.74)
$R : A \leftrightarrow B,$	$X \subseteq Y \Rightarrow R_{\Sigma}X \subseteq R_{\Sigma}Y$	$X \subseteq Y \Rightarrow R_{\Pi}X \supseteq R_{\Pi}Y$	(3.75)
$R : A \leftrightarrow B,$	$R_{\Sigma}(\bigcup_k X_k) = \bigcup_k (R_{\Sigma}X_k)$	$R_{\Pi}(\bigcup_k X_k) = \bigcap_k (R_{\Pi}X_k)$	(3.76)
$R : A \leftrightarrow B,$	$R_{\Sigma}(X_1 \cap X_2)$	$R_{\Pi}(X_1 \cap X_2)$	(3.77)
	$\subseteq (R_{\Sigma}X_1) \cap (R_{\Sigma}X_2)$	$\supseteq (R_{\Pi}X_1) \cup (R_{\Pi}X_2)$	(3.78)
$\forall\, k \in K.R_k : A \leftrightarrow B,$	$(\bigcup_k R_k)_{\Sigma}X = \bigcup_k (R_k)_{\Sigma}X$	$(\bigcap_k R_k)_{\Pi}X = \bigcap_k (R_k)_{\Pi}X$	(3.79)
$R : A \leftrightarrow B,$	$(R \cap S)_{\Sigma}X \subseteq R_{\Sigma}X \cap S_{\Sigma}X$	$(R \cup S)_{\Pi}X \supseteq R_{\Pi}X \cup S_{\Pi}X$	(3.80)
$R : A \leftrightarrow B,$	$\overline{R_{\Sigma}X} = \overline{R}_{\Pi}X$	$\overline{R_{\Pi}X} = \overline{R}_{\Sigma}X$	(3.81)

As examples of operators given:

$$\rightsquigarrow^{\sharp} \in \mathbb{P}\,\text{Propositions} \leftrightarrow \mathbb{P}(\text{Atomic Propositions})$$

then we may define operators over sets using wlp and sp by:

- $\text{wlp}(\rightsquigarrow^{\sharp}) \in \mathbb{P}\,\mathbb{P}(\text{Atomic Propositions}) \rightarrow \mathbb{P}\,\mathbb{P}\,\text{Propositions}$
- $\text{sp}(\rightsquigarrow^{\sharp}) \in \mathbb{P}\,\mathbb{P}\,\text{Propositions} \rightarrow \mathbb{P}\,\mathbb{P}(\text{Atomic Propositions})$

with the meanings:

wlp$(\rightsquigarrow^{\sharp})$: If one starts with a collection of model sets then one gets back a collection of sets that are compatible with all the model sets and not with any other model set.

sp$(\rightsquigarrow^{\sharp})$: If one starts with a collection of sets of propositional sentences then one gets back all model sets that are derivable from at least one of the starting sets.

Another interesting pair of operators can be derived from $\models$: Assignments $\leftrightarrow$ Propositions. The first operator takes a collection of assignments and gives back all the propositions that hold true under all the assignments. This is called the *theory* generated by the models. The second operator takes a collection of propositions and gives back all the models under which all the propositions hold true. This set is called the set of *models* generated by the propositions. The operators are of type:

- Theory : $\mathbb{P}$ Assignments $\rightarrow$ $\mathbb{P}$ Propositions
- Models : $\mathbb{P}$ Propositions $\rightarrow$ $\mathbb{P}$ Assignments

and defined by:

$$\text{Theory}(M) \mathrel{\widehat{=}} (\models)_{\Pi}(M)$$

$$\text{Models}(S) \mathrel{\widehat{=}} (\breve{\models})_{\Pi}(S)$$

Expressing these operators in terms of wlp and sp gives:

$$\text{Theory}(M) = \overline{\text{sp}(\overline{R})(M)}$$

$$\text{Models}(S) = \text{wlp}(\overline{R})(\overline{S})$$

3.6. Closures and Interiors

An often useful property of an operator $F : \mathbb{P}\,A \rightarrow \mathbb{P}\,B$ is the fact that it in some sense takes a set to its 'completion' or closure. An operator is said to be a closure operator when it is:

$$\begin{array}{lll} \text{monotonic:} & X \subseteq X' \Rightarrow FX \subseteq FX' & (3.82) \\ \text{idempotent:} & F(FX) = FX & (3.83) \\ \text{increasing:} & X \subseteq FX & (3.84) \end{array}$$

For example, given a relation R, the relations R^* and R^+ are a closures of R in this sense. Moreover, taking $\_^*$ and $\_^+$ as operators on relations $\_^*, \_^+ : (A \leftrightarrow B) \rightarrow (A \leftrightarrow B)$, then both can be seen to be closure operators.

Interior operators are the dual notion to closure operators. The difference is that rather than being *increasing* operators, they are *decreasing* i.e, an operator, G, obeying $GX \subseteq X$. That is an operator is an interior operation when it is:

monotonic: $X \subseteq X' \Rightarrow GX \subseteq GX'$ (3.85)
idempotent: $G(GX) = GX$ (3.86)
decreasing: $GX \subseteq X$ (3.87)

A set X is said to be closed with respect to an operator $\mathcal{O}$ if $\mathcal{O}X \subseteq X$. If X is closed with respect to $\mathcal{O}$ and $\mathcal{O}$ is monotonic then it is closed with respect to $\mathcal{O}^+$ and $\mathcal{O}^*$. For $\mathcal{O}^+$ the proof is by induction taking $\mathcal{O}$ closure as the base case. The inductive step is $\mathcal{O}^i X \subseteq X$ implies $\mathcal{O}^{i+1}X \subseteq X$ by factoring $\mathcal{O}^{i+1}X$ into $\mathcal{O}(\mathcal{O}^i X)$. By induction $\mathcal{O}^i X \subseteq X$, so, by monotonicity, $\mathcal{O}(\mathcal{O}^i X) \subseteq \mathcal{O}(X)$ and $\mathcal{O}(X) \subseteq X$. Yielding $\mathcal{O}^{i+1}X \subseteq X$. Since all $\mathcal{O}^i X$ are in X, taking the union over all i now gives $\mathcal{O}^+ X \subseteq X$. $\mathcal{O}^* X \subseteq X$ follows from adding $X \subseteq X$.

Since all operators of the form R_Σ, for relation R, are monotonic, the result is quite general.

3.7. Galois Connections and Operators of Relations

It is possible to define other operators from the operators R_Σ and R_Π by use of complement and converse operations. Interestingly, these operators often come in pairs called Galois Connections. Formally a Galois Connection is a pair of functions $(\pi_\bullet, \pi^\bullet)$ between two partial orders $(A, \sqsubseteq_A)$ and $(B, \sqsubseteq_B)$ such that

$$X \sqsubseteq_A \pi^\bullet\, Y \quad \textbf{iff} \quad \pi_\bullet\, X \sqsubseteq_B Y \tag{3.88}$$

The function $\pi_\bullet$ is referred to as the lower adjoint of the connection and $\pi^\bullet$ as the upper adjoint.

In our case the partial orders concerned are subset and superset orderings over the power sets of the domain and range of the underlying relation. The particular pair of orderings depending on the particular pair of operators involved.

Given a relation $R : A \leftrightarrow B$ there are four *forward* Galois Connections in which the lower connection is of type $\pi_\bullet : \mathbb{P}\,A \rightarrow \mathbb{P}\,B$ (and so $\pi^\bullet : \mathbb{P}\,B \rightarrow \mathbb{P}\,A$) and four *backward* Galois Connections obtained by using the converse of R so that the lower connection is of type $\pi_\bullet : \mathbb{P}\,B \rightarrow \mathbb{P}\,A$ (and so $\pi^\bullet : \mathbb{P}\,A \rightarrow \mathbb{P}\,B$). In the following chapters we are particularly concerned with the two connections, the *axiality* generated by R and the *polarity* generated by R[11]:

Name	**Upper**	**Lower**	**Law**	
Axiality	$R_\exists X \mathrel{\widehat{=}} R_\Sigma X$	$R^\forall Y \mathrel{\widehat{=}} \overline{\breve{R}_\Sigma \overline{Y}}$	$R_\exists X \subseteq Y \equiv X \subseteq R^\forall Y$	(3.89)
Polarity	$R_+ X \mathrel{\widehat{=}} R_\Pi X$	$R^+ X \mathrel{\widehat{=}} \breve{R}_\Pi X$	$R_+ X \supseteq Y \equiv X \subseteq R^+ Y$	(3.90)

The principal facts about these operators being:

[11] There is a potential for confusion between the use of R^+ for both an iterated relation and as the upper adjoint of a relation. The context and implicit typing should normally be sufficient to determine whether R^+ is a binary relation, i.e. an iterate, or an operator, i.e. an upper adjoint.

Axiality	**Polarity**
$R^\forall \circ R_\exists$ is a closure on $\mathbb{P}\,A$	$R^+ \circ R_+$ is a closure on $\mathbb{P}\,A$
$R_\exists \circ R^\forall$ is an interior operator on $\mathbb{P}\,B$	$R_+ \circ R^+$ is a closure on $\mathbb{P}\,B$
$R_\exists \circ R^\forall \circ R_\exists = R_\exists$	$R_+ \circ R^+ \circ R_+ = R_+$
$R^\forall \circ R_\exists \circ R^\forall = R^\forall$	$R^+ \circ R_+ \circ R^+ = R^+$

Some examples: First let us return to the operators $\mathrm{wlp}(R)$ and $\mathrm{sp}(R)$.

$\mathrm{wlp}(R) = R^\forall$
$\mathrm{sp}(R) = R_\exists$
$\mathrm{wlp}(R) \circ \mathrm{sp}(R)$ is a closure operation
$\mathrm{sp}(R) \circ \mathrm{wlp}(R)$ is an interior operation

Similarly we may note:

Models $= \models^+$
Theories $= \models_+$
Models $\circ$ Theories is a closure operation
Theories $\circ$ Models is a closure operation

We also note that Axialities and Polarities are intimately connected by the identities:

$$R_+X = \overline{R}_\exists\overline{X} \tag{3.91}$$

$$R^+X = \overline{\overline{R}^\forall X} \tag{3.92}$$

these identities allow us to switch back and forth between axiality's and polarities whenever the need arises.

Operators derived from binary relations and their Galois Connections often reveal interesting relations between well-known functions. Consider the membership relation of set theory $\in : A \leftrightarrow \mathbb{P}\,A$. First, define set A overlaps set B, written $\between$ by $A \between B \mathrel{\widehat{=}} A \cap B \neq \emptyset$; the collection of sets touching another by: $\mathbb{T}\,X \mathrel{\widehat{=}} \{Y \mid X \between Y\}$; the collection of sets that contain a particular element $x : A$, written $\uparrow x$, by $\uparrow x \mathrel{\widehat{=}} \{y : \mathbb{P}\,A \mid x \in y\}$; and define $\mathbb{U}\,Y$ (for $Y : \mathbb{P}\,\mathbb{P}\,A$) by $\mathbb{U}\,Y \mathrel{\widehat{=}} \{x : A \mid \uparrow x \subseteq Y\}$. Some of the operators derived from $\in$ are:

$$\in_\exists X = \mathbb{T}\,X$$

$$\in^\forall Y = \mathbb{U}\,Y$$

$$\breve{\in}_\exists Y = \bigcup Y$$

$$\breve{\in}^\forall X = \mathbb{P}\,X$$

with the Galois Connections (axialities) being:

$$\mathbb{T}\,X \subseteq Y \equiv X \subseteq \mathbb{U}\,Y$$
$$\bigcup Y \subseteq X \equiv Y \subseteq \mathbb{P}\,X$$

Likewise if we introduce the notion of the supersets of a set X of type $\mathbb{P}\,A$, written $\mathbb{S}\,X$, defined by $\mathbb{S}\,X \mathrel{\widehat{=}} \{Y : \mathbb{P}\,A \mid X \subseteq Y\}$ then the polarities give us:

$$\in_+ X = \mathbb{S}\,X$$
$$\in^+ Y = \bigcap Y$$

with the Galois Connection:

$$\mathbb{S}\,X \supseteq Y \equiv X \subseteq \bigcap Y$$

Since we have introduced the overlaps relation for the above example we also note the rather elegant equivalent definition of $R_\exists$ and $R^\forall$:

$$R_\exists X = \{y \mid X \between \breve{R}.y\} \tag{3.93}$$
$$R^\forall Y = \{x \mid R.x \subseteq Y\} \tag{3.94}$$

and of R_+ and R^+:

$$R_+ X = \{y \mid X \subseteq \breve{R}.y\} \tag{3.95}$$
$$R^+ Y = \{x \mid Y \subseteq R.x\} \tag{3.96}$$

3.8. The Use of Mathematics in Subsequent Chapters

The mathematical tools introduced in this chapter are used in the following manner in subsequent chapters:

Chapter 4: Simple set theory and the notion of propositional models are used to formalise the problem of *Knowledge on Trust*.

Chapter 5: The notions of operators, and Galois connections, derived from binary relations are used to develope a generalised form of argumentation theory. This is used in turn to define a theory of *Social Trust*. Propositions are stated in the chapter but their proofs are deferred to appendix A.

Appendix A: The proofs of propositions in chapter 5 are given. These proofs rely on various propositions about operators, Galois connections and reflexive transitive closures given in the current chapter.

4. Knowledge on Trust

In which is discussed:

- *How obtaining Knowledge on Trust is a process of building a coherent theory.*
- *A formal notion of the coherence of a theory.*
- *How we may reason about coherence.*
- *Examples of obtaining Knowledge on Trust by coherence based reasoning.*

4.1. Introduction

This chapter takes up the problem of Knowledge on Trust. The problem of Knowledge on Trust is pervasive. We gain knowledge by being told things by others. We gain information from books, newspapers and magazines, from radio and television and from the Internet. We are told inconsistent things by different people and when all accounts are in, we are left with an incomplete picture of situations. Our problem is to make sense out of all this data and arrive at a consistent account of the world we live in. To do so we must use all the information at our disposal. We have dispositions with respect to pieces of information and whether they are intrinsically believable, and we have dispositions towards sources of information and whether or not they are trustworthy. In this chapter we explore the use of this latter kind of information to help sort out a consistent account of the world. Our view of trustworthiness is rarely black and white. More usually we have some notion of the *relativity of trustworthiness*. A is more trustworthy than B but less trustworthy than C and we have no real way of, say, comparing the trustworthiness of A and D. This chapter considers how we may use the relative ordering of the trustworthiness of sources to help untangle the inconsistencies we are presented with and arrive at plausible, consistent, views of the world.

More formally, we consider the general problem of an individual, let us call them *the reasoner*, receiving information provided by multiple *sources* which are held by the reasoner to have different, possibly incomparable, degrees of trustworthiness. This relative trustworthiness is represented by a binary relation between sources that represents the idea that the reasoner holds one source more trustworthy than another. This relation is not necessarily transitive nor cycle free. The reasoner wishes to obtain conflict free, maximal, accounts of the world. That is, the reasoner wishes to incorporate as much information as possible from the sources without creating inconsistencies. In resolving information conflicts the reasoner casts out the least preferred information that leads to a conflict, using the relation between the sources of information.

Our reasoner strives for coherence in his account of the world. But what exactly does this mean?

Rather than attempt to develop a new theory of coherence, together with its justification, this chapter will examine Laurence Bonjour's model coherence [18]. This theory is then formalised as a mathematical theory and implemented as a program for coherence based reasoning. The chapter ends with three illustrative examples of the theory applied to problems of obtaining Knowledge on Trust, these are:

- Resolving conflicting evidence to arrive at a conclusion in a judicial inquiry.
- Resolving conflicting reports by counter terrorism analysts[1].
- Obtaining Knowledge of Identity on Trust within a PKI framework.

This last example serves as bridge into the next chapter.

4.2. A Theory of Coherence

Bonjour proposes that a system of beliefs is coherent if it fulfils five conditions. First a system of beliefs should be:

1. logically consistent:

   That is that it is free from classical logical contradictions.

2. probabilistically consistent:

   Probabilistic consistency says that if a reasoner is faced with an inconsistency that can be avoided by rejecting either one of the observations then the reasoner will reject the less likely of the two observations. Said differently, the reasoner has a preference for accepting the more likely observation. This is a special case of dealing with the *relative preferences over relinquishing beliefs* to resolve conflicts. In this case highly probable beliefs are relinquished less willingly than improbable beliefs. In this work the notion of probability is replaced by a weaker notion of relative plausibility which is derived from the relative trustworthiness of the sources.

   Further, Bonjour says that the coherence of a system of beliefs is:

3. increased by the presence of inferential connections between its component beliefs and increased in proportion to the number and strength of such connections.

4. diminished to the extent to which it is divided into subsystems of beliefs which are relatively unconnected to each other by inferential connections.

[1] It should be stressed that these first two examples are completely fictional.

The notion of "inferential connection" used by Bonjour is intended to capture an idea of "fitting together" but it is neither deductive entailment, which is too strong, nor consistency, which is too weak. Rather it captures a correlation between beliefs such that a reasoner prefers to believe the correlated set of beliefs whenever this is possible and does not lead to inconsistency. Let us say that the presence of a belief φ *leads to* or *supports* a belief ψ, meaning that, if φ is a belief then the reasoner prefers to also believe ψ if that will not cause an inconsistency. This is a form of non-monotonic, or default, reasoning. Correlations are represented as default rules which may be used provided that the particular use does not introduce an inconsistency. Thus these correlations sit between laws of deductive entailment and principles of extension by consistency.

Finally Bonjour requires,

5. The coherence of a system of beliefs is decreased in proportion to the presence of unexplained anomalies in the believed content of the system.

Bonjour illustrates the notion of an anomaly with an example:

> *"Suppose I am standing three feet from a pole that is four feet high. Next to my foot is a mouse, and on top of the pole is perched an owl. From these conditions I may obviously infer, using the Pythagorean theorem, that the mouse is five feet from the owl. The inference is surely adequate to justify my belief that the mouse is five feet from the owl, assuming that I am justified in believing these other propositions. And intuitively speaking, this inferential connection means that the belief that the mouse is five feet from the owl coheres with the rest of my beliefs to quite a significant extent. But none of this has any apparent connection with explanation. In particular, as Lehere points out, the inference does not in any way help to* explain *[Bonjour's emphasis] why the mouse is so close to the owl. "*

Perhaps the most significant thing here is the unstated assumption that the fact that the mouse is five feet from the owl *needs* explanation. By default we are required to assume that the mouse is aware of the presence of the owl. That the mouse prefers not to be in the presence of an owl because of its survival instincts. That "being in the presence of" is satisfied by "being within five feet of". And it is the incoherence of these propositions with our extended belief system that causes an anomaly. Viewed differently, this little story tells us that there are at least two belief systems in play, the *Pythagorean* system about geometry and measurement, and the *Animal Behaviour* system about animal behaviours and interactions. The geometrical part of the story is well supported by the Pythagorean belief system but the pertaining predator-prey relationship is in conflict with the Animal Behaviour belief system. This suggests that what is at issue is that we can obtain a logically consistent theory only at the expense of excluding some highly preferred belief, in this case the belief that mouse should not be in such an exposed position with respect to the owl.

An anomaly then arises when to arrive at a consistent set of beliefs we must exclude something that is at least as preferred as what we believe in the consistent set of beliefs. An anomaly may be removed by further information. If we consider Bonjour's story of the mouse and owl above then we might suggest the explanation that the human placed the mouse at his foot, thus placing the mouse near the owl. The naive observation that the mouse is near the owl causes an anomaly. However the elaboration that the mouse was placed there by the human provides reasons why the *Animal Behaviour* belief system may be inhibited and not come into play in reasoning about the owl and the mouse.

Below, a formal model is constructed for the fragment consisting of conditions **1**, **2** and **5**. This system will be called Preferential Coherence. It is then shown how the "correlation conditions", **3** and **4**, can be interpreted within Preferential Coherence.

An area that Bonjour does not address is what to do, if after applying all of the coherence conditions, we are still left in "two (or more) minds" i.e. if rather than being left with a single coherent system of beliefs, we are left with multiple alternative equally plausible belief systems (which necessarily must be inconsistent with one another). Rather than posit further rules to reduce such a set down to a single belief system the approach taken here adopts the approach used in both non-monotonic and paraconsistent logics and defines a process of drawing inferences from alternative collections of beliefs. Given a collection of equally plausible sets of beliefs, inferences may be drawn either sceptically or credulously. A conclusion is sceptically true if it follows from every alternative system. A conclusion is credulously true if it follows from any of the alternative belief systems. A small refinement can be made to credulous conclusions in that a credulous conclusion may be arrived at because it is true in some of the alternatives and false in others, or true in some alternatives and open (i.e. neither true nor false) in others. We will refer to these two cases, as and when the need arises, as being *determined* and *undetermined* cases (i.e. in the first case the truth value is everywhere determined by the belief systems, whereas in the latter case it is not).

The sceptical approach represents a very conservative position. We might say we have "every reason" to accept the conclusion. The credulous position, by contrast, is very liberal. We might say we have "some reason" to accept the conclusion.

4.3. The Formal Model

The model is one of propositional reasoning via coherence. In the view taken here, systems of belief are *finite* collections of statements about the world that originate from a finite collection of possible sources of information, each of which makes a finite collection of observations. Sources are ordered by a preference ordering that represents their relative trustworthiness. An observation, e.g. `Freya: ~DS & ~SLW` is a proposition, in this case `~DS & ~SLW` (asserting not `DS` and not `SLW`), labelled by the source, `Freya`, that asserts the proposition. The same proposition may be asserted by multiple sources.

Formally we will assume that the sets *Label* and *Prop* (propositions) are given and define an *observation* as a member of the set of pairs of labels and propositions, *Obs* by:

$$Obs \mathrel{\widehat{=}} Label \times Prop \tag{4.1}$$

That is, an *Obs* is a labelled proposition. Further we will assume that there is an inconsistency predicate that can tell us if a set of propositions is inconsistent, **incons** : $\mathbb{P}(Prop) \to \{0,1\}$ (note that throughout we will identify a predicate with the corresponding characteristic function).

Given a set of observations, S, we now define the projection by a set of labels L, written S_L as:

$$S_L \mathrel{\widehat{=}} \{\varphi \mid l \in L \wedge (l,\varphi) \in S\} \tag{4.2}$$

A set of labels, L, is said to be inconsistent with respect to S, written $\mathbf{incons}_S(L)$ when the set of projected propositions is inconsistent i.e. :

$$\mathbf{incons}_S(L) \mathrel{\widehat{=}} \mathbf{incons}(S_L) \tag{4.3}$$

From here we can define the set of maximally consistent subsets of labels with respect to some set of observations S.

$$\max S \mathrel{\widehat{=}} \{X \subseteq Label \mid \neg\, \mathbf{incons}_S(X) \wedge \forall\, Y \subseteq Label . X \subset Y \Rightarrow \mathbf{incons}_S(Y)\} \tag{4.4}$$

4.3.1. Consistency

Our first coherence condition is that given a set of observations, S, a coherent belief system, B, should arise from a maximally consistent set of sources:

$$\exists\, T \in \max S . B = S_T \tag{4.5}$$

We may visualise this by figure 4.1.

We next consider the two conditions, preference ordering and anomaly freeness, to potentially reduce the number of acceptable maximally consistent sets.

4.3.2. Preference

Every maximally consistent set is a potential theory. Assuming that our labels come equipped with a relation $\succ$ representing a preference ordering between sources, the preferred theories will be maximally consistent theories that do the least damage in the preference structure in the sense that holes left by omission of elements occur as far down the order relation as possible. This notion is captured by extending the relation $\succ$ over sets by:

Let us say that one set, X, is a *complement cover* of another, Y, written $X \sqsupset Y$, when the set of elements of the complement of X cover the elements of the complement of Y in the sense that:

$$X \sqsupseteq Y \mathrel{\widehat{=}} \forall\, x \in \overline{X} . \exists y \in \overline{Y} . y \succeq x \tag{4.6}$$

$$X \sqsupset Y \mathrel{\widehat{=}} X \sqsupseteq Y \wedge \neg\, Y \sqsupseteq X \tag{4.7}$$

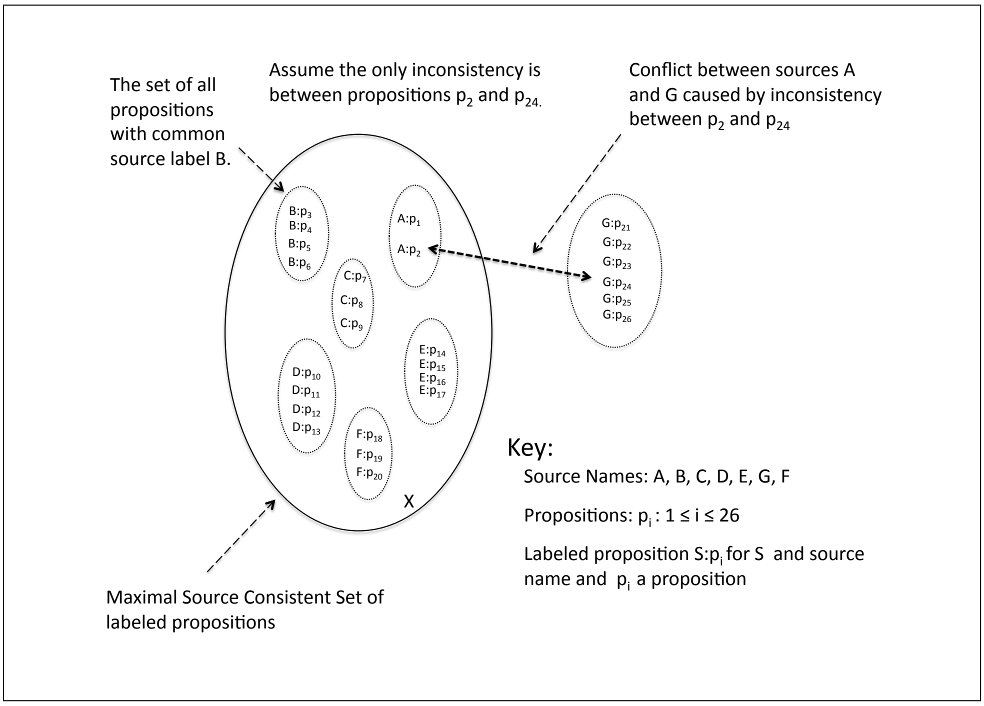
The set of all propositions with common source label B.
Assume the only inconsistency is between propositions p_2 and p_{24}.
Conflict between sources A and G caused by inconsistency between p_2 and p_{24}
$B{:}p_3$
$B{:}p_4$
$B{:}p_5$
$B{:}p_6$
$A{:}p_1$
$A{:}p_2$
$C{:}p_7$
$C{:}p_8$
$C{:}p_9$
$G{:}p_{21}$
$G{:}p_{22}$
$G{:}p_{23}$
$G{:}p_{24}$
$G{:}p_{25}$
$G{:}p_{26}$
$E{:}p_{14}$
$E{:}p_{15}$
$E{:}p_{16}$
$E{:}p_{17}$
$D{:}p_{10}$
$D{:}p_{11}$
$D{:}p_{12}$
$D{:}p_{13}$
$F{:}p_{18}$
$F{:}p_{19}$
$F{:}p_{20}$
X
Key:
Source Names: A, B, C, D, E, G, F
Propositions: $p_i : 1 \leq i \leq 26$
Labeled proposition $S{:}p_i$ for S and source name and p_i a proposition
Maximal Source Consistent Set of labeled propositions

Figure 4.1.: Maximal Source Consistency

where $x \succeq y \mathrel{\widehat{=}} x \succ y \ \vee \ x = y$.

That is, every time we find something missing from X there is something more preferred missing from Y but not vice versa.

So, for example, given the ordering $z \succ y_1 \succ x_2$, $z \succ y_2 \succ x_1$ with inconsistencies arising between x_1 and y_1 and between x_2 and y_2 the maximal set $\{z, y_1, y_2\}$ is preferred to $\{z, x_1, x_2\}$.

Then we select the $\sqsupset$-maximal sets from the maximally consistent sets of labels:

$$\max_{\succ}(S) \mathrel{\widehat{=}} \{X \in \max(S) \mid \neg\, \exists\, Y \in \max(S).\, Y \sqsupset X\} \tag{4.8}$$

4.3.3. Anomalies

Given the apparatus we have already developed, the lack of anomalies is the requirement that a maximally consistent set of beliefs does not exclude some more preferred set of beliefs. Such situations can arise with cyclic preferences.

Consider the two examples of a 2-cycle and a 3-cycle illustrated below. In the 2-cycle assume that A and B contradict one another. The maximally consistent sets are then $\{A\}$ and $\{B\}$. If we select $\{A\}$ and reject $\{B\}$ then we do not reject anything more preferred than $\{A\}$.

In the 3-cycle assume that A contradicts B, B contradicts C and C contradicts A, so again the maximally consistent sets are singletons, $\{A\}$, $\{B\}$ and $\{C\}$. Now however accepting one of them, say $\{A\}$, means rejecting a more preferred item C. Yet, of course, accepting $\{C\}$ leads to the same problem with B.

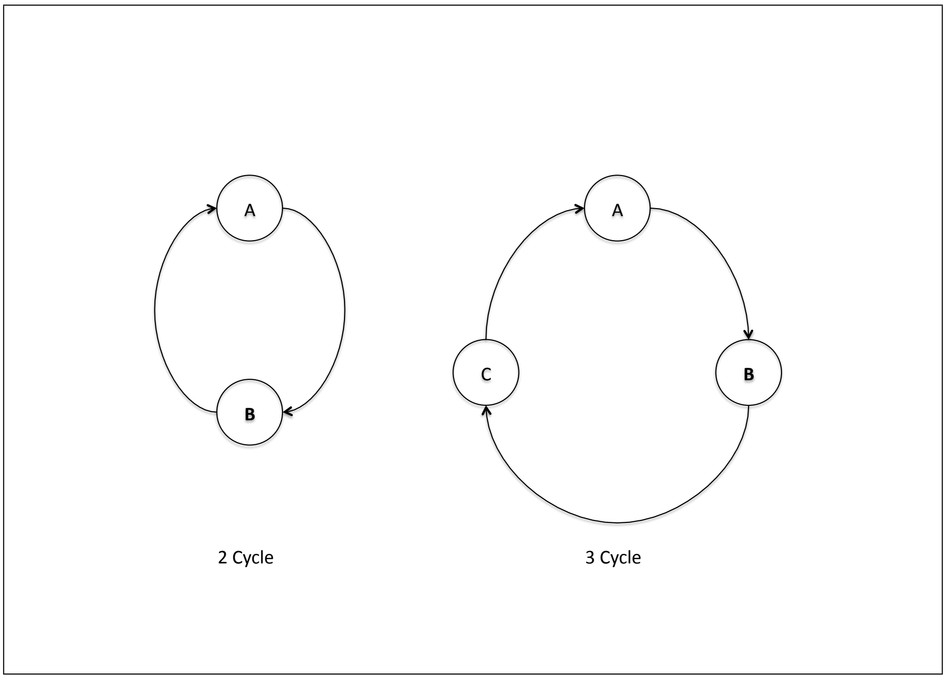

Figure 4.2.: Cycles

Such anomalies are avoided by filtering theories to exclude those which have elements outside of the theory which are more preferred than those in the theory. Specifically, we require the additional condition on theories X that $X \sqsupseteq \overline{X}$

4.3.4. Correlations

Let $\varphi \rightsquigarrow \psi$ mean that φ leads to or supports ψ. As discussed above, we wish this to have the meaning that there is a correlation between φ and ψ, such that whenever there is no reason to the contrary, the belief φ leads to the belief of ψ by the reasoner.

Let us now consider a reasoner being supplied with information from a number of external sources $S_1, \ldots, S_n$. The correlation, $\varphi \rightsquigarrow \psi$, can be represented by a labelled conditional $D : \varphi \Rightarrow \psi$ where the label D is chosen to be less than any of the external sources, i.e. $\forall i. S_i \succ D$. Note when φ is simply **true** we are left with an unconditional default belief, i.e. ψ is assumed whenever it is not contradictory to do so.

There is considerable freedom for the reasoner to place correlations anywhere in the source ordering, so the reasoner is free to regard them as less trustworthy guides than some external sources but more trustworthy than others.

4.4. A Simple Program

Overview

The theory of coherence reasoning presented above can be effectively automated. The program which does so is presented in Appendix H.

The program accepts input described by the following syntax:

Description	::=	*Preference Relation Definition* "|" *Source Assertions* hypothesis *List of Hypotheses*
Preference Relation Definition	::=	(simple | transitive) [*List of Chains*]
List of Chains	::=	*Chain* ["," *Chain*]$^+$
Chain	::=	*Source* (> *Source*)$^+$
Source Assertions	::=	*Source* : *Formula*
List of Hypotheses	::=	*Formula* ["," *Formula*]$^+$

with *Sources* being simple names and *Formula* being propositional formula with the usual operator precedence.

So, for example, consider the simple problem of trying to determine the composer of a piece of music using oneself and one's friends as sources of information. Suppose I have heard a piece of classical music played on classical guitar. I have limited knowledge of classical music but I am fairly sure it is either Bach, D. Scarlatti or S. L. Weiss. If it is Bach I do not know whether it is J.S., C.P.E. or W. F. Bach but, by default, I would assume it is C.P.E. because of some similarities with other C.P.E. works I have heard played on other instruments.

I describe the music to friends that have varying degrees of expertise in music. Let us call the friends Efran, Freya, Gabbo and Hebe with Efran > Gabbo, Freya > Gabbo, Efran > Hebe, Efran > Hebe. As I talk to them in turn my opinion on who composed the music evolves as I receive new information.

If we use the initials of the possible composers to stand for the predicate that the composer wrote the work, then the initial problem description before we have any input from our friends is:

```
transitive
Certain > Efran, Certain > Freya,

Gabbo > Default, Hebe > Default,

Efran > Gabbo, Freya > Gabbo, Efran > Hebe,  Efran > Hebe |

Certain: JSB & ~ CPEB & ~WFB & ~DS & ~ SLW  +
                ~JSB & CPEB & ~WFB & ~DS & ~ SLW  +
                 ~JSB & ~CPEB & WFB & ~DS & ~ SLW +
                 ~JSB & ~CPEB & ~WFB & DS & ~ SLW +
                 ~JSB & ~CPEB & ~WFB & ~DS & SLW ,
Default: CPEB

hypothesis
JSB,
CPEB,
WFB,
DS,
SLW
```

The keyword `transitive`, at the beginning of the problem description, means that the preference relation between sources is taken to be transitive and the transitive closure should be taken of the given preference relation (otherwise we would use the keyword `simple` to indicate that the relation is to be taken simply as specified without taking the transitive closure of given preference relation).

The sources `Certain` and `default` capture what I am certain of and my default assumption about the composer. My preference over sources is transitive and I prefer my certainty to any other information but my default is weaker than any other source of information. This results in the overall preference ordering displayed in figure 4.3. My hypotheses list is a list of simple propositions that one of the composers wrote the music. Since my default is `CPEB` and nothing stronger is known the program finds this sceptically true and, because of the exclusivity clause, that all other conclusions are sceptically false.

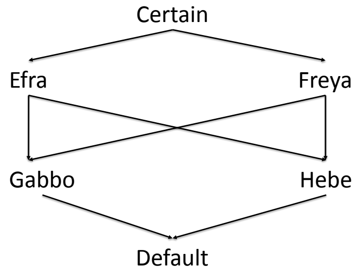

Figure 4.3.: Music Example Ordering

As outputs the program produces a summary of the input, in the form of a description of the ordering relation and the assumptions of each source, and a list of the maximal sets of sources, the “core beliefs” which are the source assertions which hold in every maximal set (if any) and the status of the hypotheses. For this example these look like:

```
Ordering
   Certain > Efran,
   Certain > Freya,
   Gabbo > Default,
   Hebe > Default,
   Efran > Gabbo,
   Freya > Gabbo,
   Efran > Hebe,
   Certain > Gabbo,
   Certain > Hebe,
   Efran > Default,
   Freya > Default,
   Certain > Default
Maximal Sets
   {Certain,Default}
Assertions
   Certain: JSB & ~CPEB & ~WFB & ~DS & ~SLW +
                ~JSB & CPEB & ~WFB & ~DS & ~SLW +
                ~JSB & ~CPEB & WFB & ~DS & ~SLW +
                ~JSB & ~CPEB & ~WFB & DS & ~SLW +
                 ~JSB & ~CPEB & ~WFB & ~DS & SLW
   Default: CPEB
Core Beliefs
   Certain: JSB & ~CPEB & ~WFB & ~DS & ~SLW +
                ~JSB & CPEB & ~WFB & ~DS & ~SLW +
                ~JSB & ~CPEB & WFB & ~DS & ~SLW +
                ~JSB & ~CPEB & ~WFB & DS & ~SLW +
                 ~JSB & ~CPEB & ~WFB & ~DS & SLW
   Default: CPEB
Hypothesis
   scep    ~JSB
   scep    CPEB
   scep    ~WFB
   scep    ~DS
   scep    ~SLW
```

The output hypotheses are formulas with a prefix indicating whether they are sceptically true, credulously true and determined or credulously true and undetermined (see 4.2 above). In this case all hypothesis have the prefix `scep`, indicating they are all sceptically true.

If we continue the example with:

- Freya saying that it is neither Scarlatti nor Weiss.
- Gabbo saying it is W.F. Bach.
- Efran saying he can rule out Weiss, and if it is Bach then it is definitely not W. F. Bach.
- And Hebe saying it is definitely J. S. Bach or Scarlatti.

then input becomes

```
transitive
Certain > Efran, Certain > Freya,

Gabbo > Default, Hebe > Default,

Efran > Gabbo, Freya > Gabbo, Efran > Hebe,  Efran > Gabbo |

Certain: JSB & ~ CPEB & ~WFB & ~DS & ~ SLW  +
                ~JSB & CPEB & ~WFB & ~DS & ~ SLW  +
                 ~JSB & ~CPEB & WFB & ~DS & ~ SLW +
                 ~JSB & ~CPEB & ~WFB & DS & ~ SLW +
                 ~JSB & ~CPEB & ~WFB & ~DS & SLW ,

Freya: ~DS & ~SLW,

Gabbo: WFB,

Efran: ~SLW & ~WFB,

Hebe: JSB + DS,

Default: CPEB

hypothesis
JSB,
CPEB,
WFB,
DS,
SLW
```

and the significant parts of the output:

```
Maximal Sets
   {Certain,Freya,Efran,Hebe}

Core Beliefs
Certain: JSB & ~ CPEB & ~WFB & ~DS & ~ SLW  +
                ~JSB & CPEB & ~WFB & ~DS & ~ SLW  +
                 ~JSB & ~CPEB & WFB & ~DS & ~ SLW +
                 ~JSB & ~CPEB & ~WFB & DS & ~ SLW +
                 ~JSB & ~CPEB & ~WFB & ~DS & SLW ,
   Freya: ~DS   &
          ~SLW
   Efran: ~SLW   &
          ~WFB
   Hebe: JSB + DS

Hypothesis
   scep    JSB
   scep    ~CPEB
   scep    ~WFB
   scep    ~DS
   scep    ~SLW
```

That is I conclude the piece is by J. S. Bach. In doing so I have discarded both my default assertion and Gabbo's assertions.

The Workings

The program is written as a functional program in Haskell. It calculates the maximally consistent sets using a simple tableaux theorem prover to derive the inconsistency of sets of theorems associated with sets of sources and then filters the maximal consistent sets of sources by the complement cover relation and the anomaly freeness condition calculated from the preference relation definition. The keyword `transitive` in the relation definition causes the transitive closure of the source preference relation to be built before it is used to produce the complement cover relation.

The truth status of the hypotheses is calculated by again using tableaux to derive the status of each hypothesis in each maximal set and labelling the output accordingly i.e.

- Sceptically true if the hypothesis follows from every maximal set.
- Sceptically false (indicated by negation of the hypothesis) if its negation follows from every maximal set.
- Credulously Determined (Credulous D) if its truth follows from some maximal sets and its falsehood follows from all others.
- Credulously Not Determined (Credulous N) if its truth follows from some maximal sets but it is not determined by some maximal sets.
- Non Determined (ND) if every maximal set fails to determine the hypothesis[2].

[2] Subsequently a more refined analysis has suggested itself. The weakness of the current analysis is that credulously true does not convey sufficient information about what is known from the analysis. An alternative analysis reflecting a more refined analysis is given by:

- Everywhere true if the hypothesis follows from every maximal set.
- Everywhere false (indicated by negation of the hypothesis) if its negation follows from every maximal set.
- Partially true if the hypothesis is either true in some sets and undetermined in all others.
- Partially false if the hypothesis is either false in some sets and undetermined in all others.
- Conflicted if there are some sets in which the hypothesis is true and others in which it is false. Note in this case there may be sets where it is undetermined, but the distinction between true

The limits on the complexity of the program is determined by the complexity of finding propositional models, and so the program is NP-complete. However this is not the whole of the complexity story. The program uses a very simple technique of model finding and proof based on propositional tableaux. Using a modern SAT solver for these tasks would vastly improve the practical scalability of the coherence reasoning process (an overview and empirical study of modern SAT solvers can be found in Sakallah and Marques-Silva [97]).

The program listing is given in appendix H.

4.5. Knowledge on Trust Reasoning Examples

We consider three examples in which Knowledge on Trust plays a central role.

The first is a judicial inquiry. In this problem the reasoner is a judge that must weigh evidence from various sources and arrive at a conclusion. The testimony is given by people of various standing, and support by various degrees of evidence.

The second is reasoning in the Intelligence Community in which the reasoner is an Intelligence Analyst trying to draw conclusions about a possible terrorist plot. The example illustrates how Preferential coherence can be used to formalise aspects of Heuer's Alternative Competing Hypothesis model of intelligence Reasoning [61].

The third example addresses the problem of reasoning about identity and, in particular, to reasoning about identity online.

4.5.1. A Judicial Inquiry

Our first example is driven by formalising Safety Cases as used in the public domain i.e. the kinds of argument that one finds taking place at public inquiries into the safety of a proposed new facility such as a Nuclear Power Station or, more mundanely, the placement of a radio mast. In such examples both experts and members of the public produce testimony in relation to the facility, its effects on the environment and its effects on people. Generally there is no clear cut *correct* decision. Rather, there are a number of possible outcomes which are determined by how the testimony is weighed by the assessor. In this case the notion of trust in the testimony is usually not a question as to whether individuals are being truthful, but rather a matter of how qualified they are to produce their opinions. Here there is a promise by the testifiers to be honest and disclosing, not only given implicitly, but given explicitly, often by oath, and under the force of law. The judgements to be made are about the respective degrees to which each witness in the proceedings is *competent* and *knowledgeable* on the subjects on which they testify[3].

Turning to the example:

or false everywhere and true, false or undetermined does not suggest a useful distinction.

- Undetermined everywhere.

[3] This is not to say that *honesty* and *candidness* may not become an issue in some proceedings, but rather by-and-large these are not normally the primary issues of such enquiries.

> *A company intends to place a radio transmitter mast in a location near a school. Legislation permits the company to operate in one of two bands, Band X and Band Y. Recent medical opinion from extensive experimentation on mice is that Band X is unsafe. An expert biophysicist believes that if Band X is unsafe then the same is true of Band Y. Although he has no explicit experimental evidence to back this up, he has great experience in assessing biological impacts of electromagnetic emissions. A technical expert is willing to testify that the company operates in Band Y.*

We might consider the ordering of plausibility of this evidence as: legislation is the most definite fact, the medical expertise and the technical expertise are on a par and that the biophysicist is expressing an opinion which is less well founded i.e. legislation > medicalExpertise, legislation > technicalExpertise, legislation > biophysicist, medicalExpertise > biophysicist, technicalExpertise > biophysicist[4] (see figure 4.4).

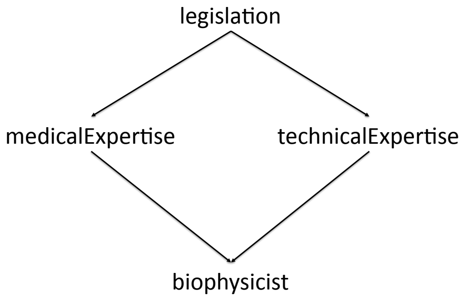

Figure 4.4.: Judicial Enquiry Ordering

The conclusion we arrive at is that it is unsafe to place the transmitter mast given the best evidence available. If now, however, additional evidence:

> *A new medical expert produces evidence to the effect that the same experiments, that were performed on mice using Band X, were performed on mice using Band Y and that there were no ill effects.*

then we would revise our conclusion, even if the biophysicist still holds the same opinion about the relation between Band X and Band Y. The reason we revise our assessment is that the new medical claim is stronger than the biophysicist's opinion because it is backed by experimental evidence.

In essence the overall safety case can be expressed as the hypothesis '`safe`' follows from the most plausible theory we can construct given the available sources of evidence and our relative evaluations of the plausibility, reliability or trustworthiness of the evidence.

Using the formalism of the Preferential Coherence Calculator, the first version of the overall safety case can be expressed as:

[4] This may be simplified by taking the relation as transitive.

```
transitive
legislation >  medicalExpertise, legislation >  technicalExpertise,
medicalExpertise > biophysicist,  technicalExpertise > biophysicist  |

legislation: (BandX & ~BandY) + (~BandX & BandY),
medicalExpertise: BandX => ~safe,
biophysicist: (BandX => ~safe) => (BandY => ~safe),
technicalExpertise: ~BandX

hypothesis
safe
```

which leads to the conclusions:

```
Maximal Sets
   {legislation
   ,medicalExpertise
   ,biophysicist
   ,technicalExpertise}

Hypothesis
   scep    ~safe
```

and the second by the same facts with the medical evidence for Band Y added:

```
transitive
legislation >  medicalExpertise, legislation >  technicalExpertise,
medicalExpertise > biophysicist,  technicalExpertise > biophysicist  |

legislation: (BandX & ~BandY) + (~BandX & BandY),
medicalExpertise: BandX => ~safe,
medicalExpertise: BandY => safe,
biophysicist: (BandX => ~safe) => (BandY => ~safe),
technicalExpertise: ~BandX

hypothesis
safe
```

which leads to the conclusions:

```
Maximal Sets
   {legislation
   ,medicalExpertise
   ,technicalExpertise}

Hypothesis
   scep    safe
```

4.5.2. Intelligence Analysis

The second example is taken from the more complex problem of reasoning about Intelligence data. This domain is more complex because, unlike the case above, we are concerned with making judgements about the *honesty* and *candidness* of our sources, as well as their degree of *competence* and *knowledge*.

It is an attempt to formalise part of a proposed systematic approach to Intelligence data analysis set forth by Richard Heuer in his book "The Psychology of Intelligence Analysis" [61]. Heuer, as the book title suggests, is primarily concerned with the psychology of Intelligence Analysts and how analysts can become trapped into seeking evidence to support preconceived conclusions rather than finding the best conclusion that fits the facts available. As part of his solution to the problem Heuer sets out an approach to Intelligence Analysis based on comparing Alternative Competing Hypotheses (ACH). Jack Davis, in the introduction to Heuer's book [61, page xxiii], describes ACH succinctly:

> *At the core of ACH is the notion of competition among a series of plausible hypotheses to see which ones survive a gauntlet of testing for compatibility with available information. The surviving hypotheses - those that have not been disproved - are subjected to further testing. ACH, Heuer concedes, will not always yield the right answer. But it can help analysts overcome the cognitive limitations discussed in his book.*

For our purposes we may split Heuer's analysis into hypothesis formation, which inherently relies on the analyst's experience to form "good" alternative hypotheses, and the evaluation of the consistency of those hypotheses with various reports. It is the latter part of the process for which we use coherence based reasoning. Heuer uses an informal logic for performing this process. Other authors have formalised Heuer's approach in a probabilistic setting. The problem with this is that whereas certain sources of information have well ascribed probabilities, other sources do not and the assignment of probabilities is performed to rank the relative credence one should place in the sources. This forces a total order on sources that may be inappropriate. The alternative taken here is the use of transitive binary relation to "order" the sources by relative trustworthiness.

The essence of the proposal here is that sources of information, including the agent's own insights and experience, can be ordered according to trustworthiness, and the maximum amount of information should be used that is compatible with the avoidance of contradictions. In the case of contradictions, information should be discarded in a manner that does *least damage* to the preference ordering.

As an example, we consider a situation in which an analyst has various sources of information about a possible terrorist attack. Firstly, she has specific knowledge of attacks acquired over the years which places bounds on her expectations of the terrorist activities. This knowledge is incomplete and may be incorrect but it is the most certain knowledge she has. Then there is information that she may obtain from sources such as telephone intercepts, police reports, sightings at transit points and informers. This information has various degrees of credibility and can be arranged in a hierarchy. The general rule is that she will consider all information as true unless a conflict arises between sources. In such cases she will discard the less credible sources involved in the conflict in order to gain consistency. Note that this will mean a single conflict causes all the information provided by the least credible sources to be disregarded. If this is an issue for a particular source, e.g. the source is regarded as more credible for certain types of information than for others, this is handled by splitting the source into multiple independent sources, each providing a part of the source's information.

Our analyst has a set of *background* assumptions built up from experience:

- That potential terrorists fall into two categories, Professional and Amateur. Professionals are further divided into Career terrorists and Disposable terrorists. Career terrorists carry out repeated acts of terrorism whereas Disposable terrorists carry out suicide missions. Professionals tend to operate with support teams and the identification of the presence of a support team is sufficient to indicate an attack will be professional.
- The modes of attack that are available are Sniper, Bomb and Mortar, with Bomb divided into Placeable, Car Bomb and Suicide Bomb. A career terrorist may attack as a Sniper, with a Placeable Bomb or with a Mortar.
- A Disposable terrorist will attack with a Suicide Bomb or a Car Bomb (which is often regarded as a form of Suicide Bomb). And an Amateur will act as a Sniper or use a Suicide Bomb.
- The only true distance attack option is the Mortar and this is only used by Career terrorists.
- Amateurs will use homemade explosives, whereas professionals will steal or purchase explosives, or purchase explosive precursors (not necessarily legally).
- The quantity of materials involved will indicate the likely size of an explosive device. So reports of theft or purchase of small quantities of explosive will indicate Placeable or Suicide Bombs, whereas large quantities will indicate Car Bombs. Purchase of large quantities of precursors will indicate a Car Bomb but purchase of small quantities of precursors is likely to go unnoticed and also is unlikely to be indicative (since by their very nature they can be purchased for legitimate ends).
- Career terrorists always have an escape plan, Amateur and Disposable ones do not need one.

For this particular example we will assume that the analyst believes there will only be a single attack (as opposed to the possibility of multiple simultaneous attacks).

In addition to these background assumptions, the analyst also has some weak *default assumptions* that, given no evidence to the contrary, the analyst takes as good working hypotheses. In this case:

- The attack will be carried out by an amateur.

The analyst partially orders the information sources by reliability and trustworthiness. In doing so she must also assess where her own assumptions sit with respect to other sources of information. For example, she may regard intercepts and police reports as definite pieces of information that are more reliable than her own assumptions, whereas she may regard sightings as less reliable, and informers as simply less trustworthy than her own background assumptions (but as a more reliable guide than her default assumptions). In examining her own background assumptions she might find that they fit into logical groupings such that, if any item in the grouping was to be contradicted by a preferred source, then she would give up the entire group of assumptions. In this example we divide the analyst's background assumptions up into assumptions about the weapons that might be used in an attack (WEAPONS), assumptions about the type of attacker (ATTACKER_TYPE), assumptions about the preferred weapons associated with a type of attacker (ATTACKER_PREFERENCES) and assumptions about whether there is a single or multiple attack (SINGLE_ATTACK) and, if the attack uses explosives, whether or not there is a single type of bomb involved (SINGLE_BOMB_TYPE). The preference ordering is illustrated in figure 4.5 (there may also be a set of definitions which is taken as trustworthy, one might say, by definition).

The analyst also forms a collection of hypotheses of interest, in this case:

- Type of attacker:
  - Amateur,
  - Professional:
    - Career,
    - Disposable
- Type of attack:
  - Bomb:
    - Placeable Bomb,
    - Suicide Bomb,
    - Car Bomb
  - Sniper,
  - Mortar

The Preferential Coherence model of the situation is:

```
transitive

DEF > INTERCEPT,
DEF > POLICE,
INTERCEPT  > RULESW,
INTERCEPT>  RULEST,
INTERCEPT > RULESP,
INTERCEPT > RULESE1,
INTERCEPT > RULESB,
```

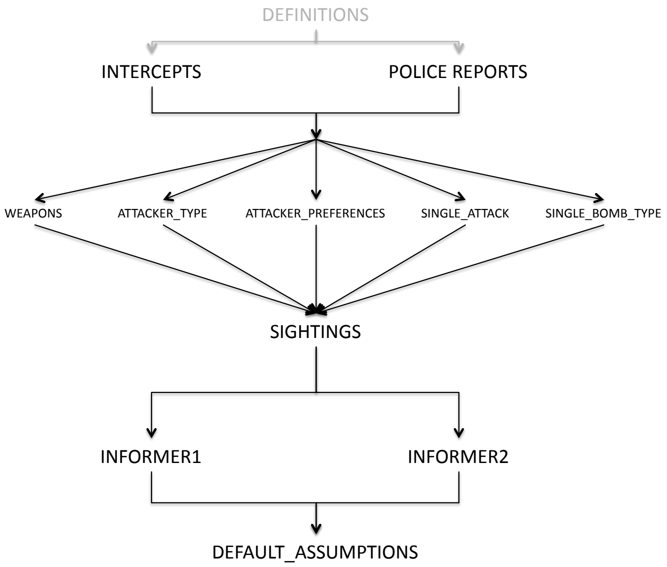

Figure 4.5.: Information Source Ordering

```
INTERCEPT > RULESE2,

POLICE  > RULESW,
POLICE >  RULEST,
POLICE > RULESP,
POLICE > RULESE1,
POLICE > RULESB,
POLICE > RULESE2,

RULESW > SIGHTINGS,
RULEST > SIGHTINGS,
RULESP > SIGHTINGS,
RULESE1 > SIGHTINGS,
RULESB > SIGHTINGS,
RULESE2 > SIGHTINGS,

SIGHTINGS > INFORMER1,
SIGHTINGS > INFORMER2,
INFORMER1 > ASSUMPTIONS,
INFORMER2 > ASSUMPTIONS |


DEF: Professional <=> Career + Disposable,
DEF:  BombThreat <=> PlaceableBomb + CarBomb + SuicideBomb,
RULESW: QuantityTheftExplosive + QuantityPurchasedPrecursors
                                    => CarBomb,
RULESW: SmuggledRifle + PurchasedRifle => Sniper,
RULESW: DistanceAttack => Mortar,
RULESW: SmallTheftExplosive + SmallPurchaseExplosive
                         => PlaceableBomb + SuicideBomb,
RULESW: QuantityTheftExplosive + SmallTheftExplosive
                                    => Professional,

RULEST: Support => Professional,
RULEST: QuantityPurchasedPrecursors => Career,
RULEST: EscapePlan => Career,
RULEST: ~ EscapePlan => Disposable + Amateur,

RULESP: Disposable  => SuicideBomb + CarBomb,
RULESP: Career => PlaceableBomb + Mortar,
RULESP: Amateur => SuicideBomb + CarBomb + Sniper,
RULESP: Mortar => Professional,

RULESE1: (Career & ~ Disposable & ~ Amateur) +
         (~ Career & Disposable & ~ Amateur) +
         (~ Career & ~ Disposable &  Amateur),
RULESB: BombThreat =>
         (PlaceableBomb & ~CarBomb & ~SuicideBomb) +
         (~ PlaceableBomb & CarBomb & ~SuicideBomb) +
         (~ PlaceableBomb & ~CarBomb &  SuicideBomb),
RULESE2: (Sniper & ~ Mortar & ~ BombThreat) +
         (~ Sniper & Mortar & ~BombThreat) +
         (~ Sniper & ~ Mortar &  BombThreat),

ASSUMPTIONS: Amateur,

INTERCEPT: BombThreat,

SIGHTINGS: Support,

POLICE: SmallTheftExplosive,

INFORMER1: Sniper,

INFORMER2: ~ EscapePlan,

INTERCEPT: EscapePlan,

INTERCEPT: PlaceableBomb & Sniper
```

```
hypothesis

Amateur,
Professional,
Career,
Disposable,
PlaceableBomb,
SuicideBomb,
Sniper,
CarBomb,
Mortar
```

The conclusion obtained from this input is

```
Maximal Sets
   {DEF, RULESW, RULEST, RULESP, RULESE1,
    RULESB, INTERCEPT, SIGHTINGS, POLICE
    ,INFORMER1}

Hypothesis
   scep    ~Amateur
   scep    Professional
   scep    Career
   scep    ~Disposable
   scep    PlaceableBomb
   scep    ~SuicideBomb
   scep    Sniper
   scep    ~CarBomb
   ND      Mortar
```

That is, the analyst concludes that the information sceptically supports the hypothesis that she is dealing with a career terrorist, who will mount an attack by using a placeable bomb and combined with a sniper attack. The analysis leaves the question of a mortar attack undetermined.

The example, as presented here, is static. A dynamic version is presented in the author's paper "Boolean Coherence and the ACH method" [53], in which information is acquired dynamically and the best estimate of the situation changes non-monotonically in response to each additional piece of information.

In the dynamic version the source assertions are added incrementally with a step in the calculation being a source making a report. The result is a changing pattern of assessments with hypotheses switching between confirmed (sceptically true), denied (sceptically false) and undetermined as each new update occurs. The resulting output table is reproduced here as table 4.1. This table highlights the difficulty of taking actions in many situations. Not only must one choose an action that is appropriate for the situation as currently perceived, but one must choose an action that minimises one's regret should that perception be in error due to lack of crucial information.

TIME	1	2	3	4	5	6	7
SOURCE	INTERCEPT	SIGHTINGS	POLICE REPORTS	INFORMER 1	INFORMER 2	INTERCEPT	INTERCEPT
REPORT	Bomb Threat	Support	Small Theft Explosive	Sniper	No Escape Plan	Escape Plan	also Sniper
Amateur	+	-	-	-	-	-	-
Professional	-	+	+	+	+	+	+
Career	-	+/-	+/-	+/-	-	+	+
Disposable	-	+/-	+/-	+/-	+	-	-
Bomb Threat	+	+	+	+	+	+	+
Placeable Bomb	-	+/-	+/-	+/-	-	+	+
Suicide Bomb	+/-	+/-	+/-	+/-	+	-	-
Car Bomb	+/-	+/-	-	-	-	-	-
Sniper	-	-	-	-	-	-	+
Mortar	-	-	-	-	-	-	+/-

Table 4.1.: Evolution of Hypotheses

4.5.3. Reasoning about Identity and Certificate Authorities

The problem of finding sufficient grounds for establishing identity has a long pedigree in both fact and fiction. For example, the perennial plot of the long lost heir that appears to claim an inheritance, or the sixteenth century story of Martin Guerre[5].

[5] The basic story of Martin Guerre is well known and has been dramatised in books, film and on stage, as well as being document in historical accounts. The basic outline of the story is:

1538 Aged 14, Martin married Bertrande de Rols.

1546 After 8 years Bertrande bears Martin their first child, a son.

1548 Martin disappears after his father accuses him of theft.

1556 A man claiming to be Martin returns to Bertrande. Martin's farther is dead and Bertrande, Martin's uncle Pierre and Martin's 4 sisters all accept the man as Martin.'Martin' lives with Bertrande for 3 years. Bertrande bears him 2 children. He also claims his inheritance from his father's estate

1559 A soldier passing through the village claims that 'Martin' is an impostor and the real Martin lost a leg during the wars. Pierre and his sons-in-law attack 'Martin' but Bertrande intervenes. A trial ensues in which 'Martin' is accused of being an impostor. He is acquitted.

1560 Pierre discovers witnesses as to the true identity of the impostor. A second trial ensues and this time 'Martin' loses and is sentenced to death. 'Martin' appeals against the verdict. The appeal is going well and the appeal judge favours 'Martin's' account that Pierre has bribed witnesses to perjure themselves. At this point a man with a wooden leg appears and claims to be the real Martin Guerre. Pierre, Bertrande and Martin's 4 sisters agree that the newcomer is the real Martin Guerre and the impostor is found guilty. In September 1560 the impostor is sentenced to death. At some point after sentencing the impostor confesses and explains how one day he was mistaken for the long vanished Martin by two people and how these two helped him learn details of Martin's life.

Once continuity of presence is lost and time has passed, establishing identity rapidly becomes difficult. Taking the problem to a further level, is the novel "WHO" by Algis Budrys. Set against a cold war background, Budrys considers what happens when both continuity of contact and the identifying features of an individual are lost. All that is left is what the individual can recount and, to some extent, what the individual can do. In the novel the central character, cannot re-establish his identity, and is forced to build a new life and a new identity.

But all this begs the question as to what we mean by identity?

Our interest in identity is not an interest in identity for itself. Rather, we are interested in identity because it gives us access to other information about people. If I know the identity of someone, then I may know, perhaps because I have been told by others, where they live, or what their credit rating is, or whether, or not, they are honest. In this sense, identity is quite a practical matter. Our interest is in the information that we can derive from identity and in the use of that information to make various judgements about others[6].

The practical problem of identity is that of tying some token, or some *indicator of identity*, to some information the token is intended to guarantee. Thus practical identity is not a fixed notion but varies according to context. For example, from the point of issuing an ATM card, the only real concern of the bank is that you are the account holder. From the point of view of using an ATM, the only concern of the bank is that the right combination of card and PIN have been presented, there being a contract in place between you and the bank that the PIN has not been disclosed to others and so can stand in lieu of your identity. From the point of view of creating an account, the bank is required, by law, to use a stronger notion of identity, usually multiple pieces of documentary proof of identity, as part of governmental measures against money laundering and organised crime.

The general pattern of establishing identities is illustrated by the ATM example. An identity is established by carrying out some checks. Some token of identity is then established and this token can be used in place of performing the checks on future occasions. As the tokens stand in lieu of checks they may themselves be used as part of the process of establishing further tokens for other contexts. This includes using multiple tokens to cross validate identity before a new token is established, as required, for example, to open a bank account[7].

Tokens themselves are of two forms. They may be *natural tokens* which exist in some *natural correspondence* to the property concerned so that association, once established, is undeniable, or unfakable. In this case the issue is one of the safety of the recording of the association. Alternatively it may be an *artificial token* which exists in some *artificially maintained correspondence* with the property, which is to say it is a matter of convention, in which case it is underwritten by some mechanism that maintains the convention. At issue here is both the safety of the recording of

[6] Identity is therefore a social phenomenon. Identity exists only in relation to others and tokens of identity carry the meaning of one group guaranteeing some property or behaviour of an individual to another group.

[7] Note that the notion of token here includes such measures as a person recognising a face, the scanning of a fingerprint, the taking of a DNA sample, as well as the issuing of a card, documents, PIN, digital certificate or password.

the correspondence and the safety of the maintenance of the convention[8].

As the web of tokens around an individual becomes dense the individual's position in society becomes more defined and the individual finds it increasingly easy to establish yet more tokens for different purposes. This rich web of tokens corresponds to a rich set of abstract mechanisms for trust, each underwritten by the processes of establishing the token and maintaining the token's validity over time.

The exemplar of identity in the online world is Public Key Infrastructure (PKI). Much has been written about PKI from the point of view of cryptography and security. And the exemplar of its failure is that of DigiNotar as discussed in the thesis introduction. Here we address the trust aspects of PKI and see how local failures undermine trust in identities. Since we are concerned with the trust model rather than engineering we may elide many technical details and describe matters simply.

The security of the PKI model is based on the idea that it is possible to pass documents around in some manner, such that their contents are private, that it is not possible to tamper with their contents, and that the documents can be signed in such a way that the signature is unforgeable and uniquely identifies the signer.

Within this framework it is possible for an individual to produce a document to attest to the identity of the signer of another document. Generally documents are used to make claims about the trustworthiness of individuals for some purpose or role. We will elide this latter detail in our treatment since, although it is of practical importance, it adds little to the discussion of identity on trust. We will call a signed document making an identity claim a certificate. More specifically a certificate makes a conditional identity claim of the form *if* A *is Trustworthy for identity then* B *is Trustworthy for identity* where A is the signer of the certificate and B is some other individual. So the certificate is essentially a message from the signer saying, if you trust in *my* identity and you trust *me* to make claims about the identity of others then you can trust this *other*. Clearly it is possible to build chains of these certificates, leading from trust in the signer at the beginning of the chain to trust in an individual at the end of the chain, provided, of course, we have trust in the intermediaries to make claims about the identity of others.

Here we see a pattern. We trust someone because there is a believable claim that they are trustworthy and we wish to trust them, as captured by the conditional assertion on our part that trustworthiness implies trust[9]. The trustworthiness assertions are generally also conditional since, to believe them, we need to trust the source of the assertion. So there must be some licensing assertion to convert the conditional assertion into a simple fact.

A typical real world situation is where an organisation takes on the role of a Public Certificate Authority which will issue certificates for the identity of some Company Certificate authority, which will in turn issue certificates for the identity of the Company Websites. Reasoning about the validity of a website identity then

[8] Traditional systems and cryptographic security are concerned with the mechanisms by which tokens are bound to the properties and how the binding is recorded. Economic and game theoretic views of security can be seen as addressing aspects of how the conventions are maintained.

[9] Without this assertion on our part then an individual can be as trustworthy as we like without us actually trusting them.

has the following pattern:

- I am willing to trust the Company Website if the Company Certificate Authority has issued a certificate for the website and I trust the Company Certificate Authority.
- I am willing to trust the Company Certificate Authority if a Public Certificate Authority has issued a certificate for the Company Certificate Authority and I trust the Public Certificate Authority.
- The Company Certificate Authority has issued a certificate for the website.
- Generally, unless there are reasons to the contrary, I am willing to believe the company issues its certificates responsibly.
- A Public Certificate Authority has issued a certificate for the Company Certificate Authority.
- A licensing body says the Public Certificate Authority is trustworthy in the matter of issuing company certificates
- Generally, unless there are reasons to the contrary, the assertions of the licensing body about trustworthiness are to be believed.

Let `License` be the licensing body `CA` be the public certificate authority, `CompanyAuth` be the company certificate authority, `CompanyWeb` be the company web site and let

- `TrustCompanyWeb` stand for "I trust the company website".
- `TrustworthyCompanyWeb` stand for "The company website is trustworthy".
- `TrustworthyCompanyAuth` stand for "The company certificate authority is trustworthy".
- `TrustworthyCA` stand for "The public certificate authority is trustworthy".

```
simple |

Me: TrustworthyCompanyWeb => TrustCompanyWeb,
CompanyAuth: TrustworthyCompanyAuth => TrustworthyCompanyWeb,
CA: TrustworthyCA => TrustworthyCompanyAuth,
License: TrustworthyCA

hypothesis
TrustCompanyWeb
```

From which we draw the conclusion that the hypothesis is sceptically true and we can sceptically trust the company website. If now we see a story in a newspaper to the effect that `CA` is untrustworthy because of a failure of internal procedures i.e. adding the condition:

```
News: ~ TrustworthyCA
```

then the hypothesis becomes non-defined, i.e we cannot conclude trust nor lack of trust in the website, because the claim of the licensing body has been undermined by `News`. Indeed there are two possible interpretations of the world, one in which `News` is correct and one in which `License` is correct. If we order the information sources `News` and `License`, with `News` > `License`, then we can conclude that `CA` is not trustworthy, but this still means that we can only conclude that we cannot determine whether or not to trust the company website. However if we had the preference order the other way around i.e. `License` > `News`, then the trust in the licensing body would not be undermined and we could establish sceptical trust in the website.

In this case we have modelled the idea that the newspaper attacks the claims of the licensing body that CA is trustworthy. But the paper might also attack the claims of the CA directly by saying the certificates content cannot be trusted i.e. the claim `TrustworthyCA => TrustworthyCompanyAuth` is untrustworthy. Again trust in the website collapses as the chain of inferences supporting that trustworthiness of the website is undermined.

Elaborating the example to consider multiple Public Certificate Authorities allows us to examine issues around cross signing between Public Authorities[10].

In this case we will assume two Public Certificate Authorities, two Company Certificate Authority and two Company Websites.

```
simple |

Me: TrustworthyCompanyWeb1 => TrustCompanyWeb1,
CompanyAuth1: TrustworthyCompanyAuth1 => TrustworthyCompanyWeb1,
CA1: TrustworthyCA1 => TrustworthyCompanyAuth1,
License: TrustworthyCA1,

Me: TrustworthyCompanyWeb2 => TrustCompanyWeb2,
CompanyAuth2: TrustworthyCompanyAuth2 => TrustworthyCompanyWeb2,
CA2: TrustworthyCA2 => TrustworthyCompanyAuth2,
License: TrustworthyCA2

hypothesis
TrustCompanyWeb1,
TrustCompanyWeb2
```

In the first instance we model the situation in which I wish to visit the websites of two different companies, we regard the licensing body as licensing trust in both

[10] The original 1978 certificate PKI model of Loren Kohnfelder [63] envisaged a single "root" Public Certificate Authority at the start of all certificate chains. When certificate systems were implemented commercially this did not happen and many distinct Public Certificate Authorities emerged. The practical problem then arose of trusting certificates signed by different authorities leading to a proliferation of root certificates on computers. To ease the problem authorities adopted a cross signing approach where each authority asserted the trustworthiness of the other authorities. So the chain up to an authority could be continued by another authority.

Public Certificate Authorities and there is no cross signing. If nothing else is added then I trust both Company Websites.

If we prefer News over License and the source News undermines the licensing assertion, TrustworthyCA1, made by License, i.e.

```
News: ~ TrustworthyCA1
```

then all the assertions of the licensing body are regarded as suspect. As a result it is not possible to establish trust in either website.

Now this may be an appropriate reaction when, for example, we first hear of the shortcomings of a CA's behaviour. The licensing body did not detect a particular CA's failures, so why should it be trusted to have made the right calls in other cases?

If, however, the reporting had been different and News had reported that the assertion of CA1 was false:

```
News: ~ (TrustworthyCA1 => TrustworthyCompanyAuth1)
```

(with News > CA1), then trust is lost in the Website of Company1 but not in the Website of Company2. This might arise, for example, because a weakness in the cryptography used by CA1 leads to it being possible for a third party to forge certificates for CA1. Moreover any other sites whose trustworthiness depends on CA1 also lose trust (that is unless there is an independent route to establishing their trustworthiness).

Returning to the undermining of License, suppose that there are two licensing bodies, or two independently trusted acts of licensing by the same body that can be viewed as independent so that failures of one do not contaminate the other. Further, we have each CA independently licensed, i.e.

```
simple News > License1 |

Me: TrustworthyCompanyWeb1 => TrustCompanyWeb1,
CompanyAuth1: TrustworthyCompanyAuth1 => TrustworthyCompanyWeb1,
CA1: TrustworthyCA1 => TrustworthyCompanyAuth1,
License1: TrustworthyCA1,

Me: TrustworthyCompanyWeb2 => TrustCompanyWeb2,
CompanyAuth2: TrustworthyCompanyAuth2 => TrustworthyCompanyWeb2,
CA2: TrustworthyCA2 => TrustworthyCompanyAuth2,
License2: TrustworthyCA2,



News: ~ (TrustworthyCA1)


hypothesis
TrustworthyCA1,
TrustworthyCA2,
TrustCompanyWeb1,
TrustCompanyWeb2
```

Clearly if `News` undermines `CA1` this has no effect on `CA2`. But if we add cross-signing:

```
CA1: TrustworthyCA1 => TrustworthyCA2,
CA2: TrustworthyCA2 => TrustworthyCA1,
```

then we see that `CA2`'s assertion leads to a contradiction which undermines `CA2` leading to neither website being regarded as trusted. So with cross-signing the trustworthiness of a collection of certificate authorities becomes susceptible to an attack on one of them. That is, because we know one claim of trustworthiness that ensues from combination of licensing and cross-signing is false then all similar claims are suspect.

Also of note is that the status of the trustworthiness of `CA1` and `CA2` depends on how `News` stands in relation to `License1` and `License2`. If `News` is preferred over both CA's then both CA's are sceptically not trustworthy. But if either licensing body is preferred over `News` then both CA's are sceptically trustworthy (and the websites trusted). Any other arrangement leads to a lack of determination of the trustworthiness of the CA's.

In this example we have become less concerned with the *content* provided by sources and more concerned with how sources *undermine*, or *attack*, one another's positions. In the next chapter we will turn to considering systems in which the content is stripped away, or rather specified implicitly, and the notion of *attack*, and also *support* are studied in their own right.

4.6. Conclusions

This chapter has introduced Bonjour's theory of coherence and formalised it as a mathematical theory and discussed its implementation as a program. It has illustrated how this theory may be used to reason in three very different example domains where the problem of Knowledge on Trust arises. The program has then been used in three examples to calculate a plausible view of the state of the world. The final example illustrates how knowledge on trust connects the notions of trust, trustworthiness and the licensed belief in both.

5. Social Trust and Argumentation Theory

In which is discussed:

- *How social trust can be modelled by Argumentation Systems.*
- *How Argumentation Theory may be formalised via Galois Connections.*
- *How argumentation systems may be evaluated using model checking.*
- *And finally how privacy is an application of social trust.*

5.1. Introduction

Trust acts to remove the uncertainty in social transactions. Indirect trust does this by using opinions gathered from the social network in which we are embedded. Indirect trust acts in lieu of classical trust allowing us to form initial trust impressions of individuals socially remote in our network. This allows social life to continue without the extensive costs of insurance and hedging of every potentially risky transaction.

The essence of the problem considered here is: given I do not have personal, direct and specific evidence that someone, say Sam, is trustworthy, how am I going to establish enough initial trust to interact with him? Suppose Sam is a butcher I have not used previously. On what basis might I gain initial trust in him to buy my meat from him? Living as I do in a small village, I would consult other villagers as to Sam's reliability and trustworthiness as a butcher. In doing this I would be looking for recommendations but, those recommendations may not be direct. My neighbours are vegetarian, although tolerant of omnivores, such as myself. Although they cannot recommend Sam directly, they can tell me that Mrs. Johnston swears by Sam's prime cuts. Other sources, directly and indirectly, tell me that Sam is honest and straightforward, keeps clean premises and regularly wins awards for food quality and hygiene. In assessing this testimony I assess whether, or not, I have reason to doubt any of it. Are my informants reliable? Is their testimony coherent overall? If there are detractors from Sam's good name as a butcher, are the detractors' opinions worth listening to or do my informants as a body rebut such negative appraisals?

What seems to be going on in such an assessment is that I am picking out a particular community or club that presents a coherent position on Sam as a butcher. The members of this club implicitly trust and agree with one another in that they do not contradict one another and overall they do not rebut other views. Moreover both I and Sam are members of this club and there is an explicit path of trustworthy recommendations that connect me to Sam. If it is possible to find such a club and such an explicit path then I am willing to trust Sam (at least until I get direct evidence to the contrary through personal experience) and try him as my meat supplier. We might say that the essence of this approach to initial trust is that Sam and I are socially connected within a societal matrix in which I have implicit trust. The problem of gaining initial trust is determining the existence of an implicitly trusted social matrix that supports trust referral between myself and other individuals.

The club that forms the implicitly trusted social matrix should obey some "reasonableness conditions" which ensure some degree of consistency of outlook between the club members. It should be composed of individuals which do not distrust one another. The club overall should defend itself against outside distrust in that, if an outsider distrusts some member of the group, then some member of the group should provide reasons for distrusting the outsider's opinion. Moreover, no one in the group should trust an outsider that distrusts a group member (for if they did then they trust the opinion that someone in the group is untrustworthy). More generally still, the opinions of members of the group about outsiders should have some level of agreement e.g. they should not be in direct conflict with one another, so that one club member trusts an outsider and another distrusts them. Generally clubs will vary in which of these external conditions they adopt.

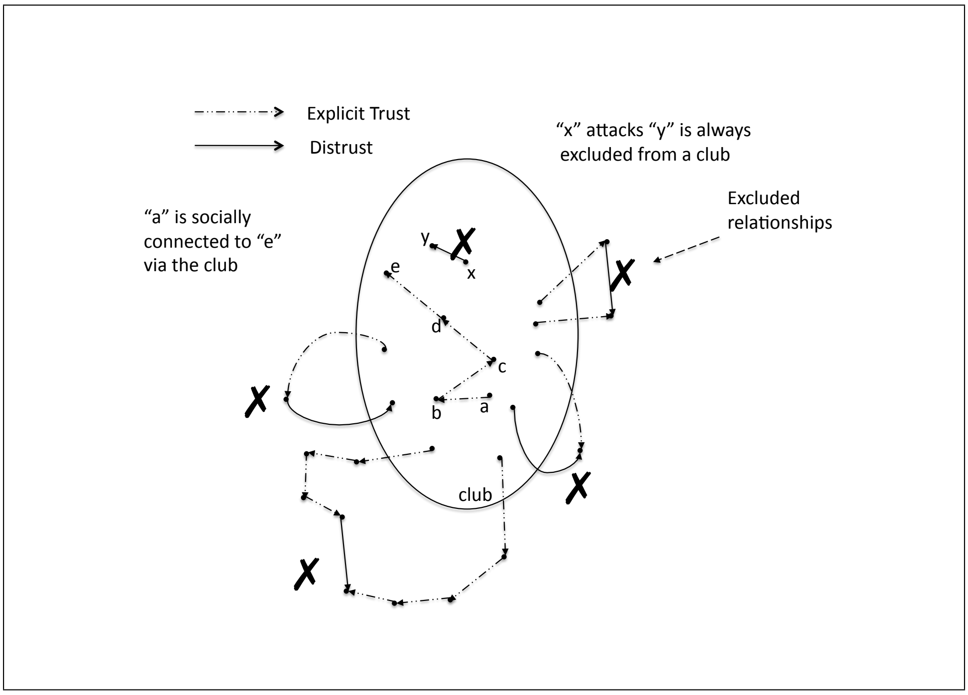
Explicit Trust
Distrust
"x" attacks "y" is always excluded from a club
Excluded relationships
"a" is socially connected to "e" via the club
y
x
e
d
c
b
a
club

Figure 5.1.: Examples of Club Conditions

The model developed in this chapter is built from two relations that can directly connect individuals, direct trust and direct distrust.

This gives rise to four possible directed relationships between any pair of individuals, *A* and *B*.

- *A* may trust *B*.

- A may distrust B.
- A may be neutral towards B by neither trusting nor distrusting B.
- A may be paradoxical towards B by both trusting and distrusting B.

Individuals may give referrals to one another in the context of some particular enquiry. By giving a referral, an individual transfers trust, for a particular topic, to another. Referrals may be either positive or negative. A positive referral means that there is reason to trust the referee, and a negative referral means that there is reason to distrust the referee. The most basic logical process with referrals is the cancellation effect of negative referrals i.e. if person A tells you that person B is to be distrusted and person C tell you person A is to be distrusted, then you may decide that you distrust person A's distrust of person B (if you accept C's testimony).

The process of giving referrals extends these relations from direct to indirect connections, leading to the development of indirect trust between individuals. Indirect trust is here regarded as a trust stance. If I have indirect trust of another individual then I am biased towards trusting them. However, this stance may be overturned by direct evidence to the contrary.

From the point of view of trust, our first consideration is defining which groups of individuals have no reason to distrust one another. If we assume that individuals have a trusting stance we may equally describe this as the group implicitly trusting one another. If we first consider the distrust relationship (negative referrals) then we want groups in which no member distrusts another member. A stronger condition on the group is that it defends itself against outside attacks on the trustworthiness of its members by attacking the trustworthiness of anyone outside the group that attacks the trustworthiness of a member of the group. Such defence does not require that the entire group be able to rebut the attack but merely that some members within the group can rebut the attack. The abstract form of such structures i.e. of a set, referred to as a set of arguments, with binary relations defining attacks and supports between members of the set and the extraction of consistent and self-defending subsets, has been studied as the subject of Argumentation Theory, a subject to which we now turn before resuming our consideration of the process of gaining initial trust.

The following sections develop an account of argumentation theory and develop it via the use of Galois Connections. Where it is useful to recall a definition or equation from chapter 3, a back reference, of the form $\_^{[\text{equation number}]}$, is given in the text[1]. A novel approach is then developed to the evaluation of finite argumentation systems by compiling the argumentation systems into Boolean Networks that can be evaluated by model checking using standard model checking tools[2].

Argumentation theory is presented first as the standard formulation of Dung [17, 34], which considers a single relation of attack (or in our interpretation *distrust*)

[1] Few such links are required in this chapter since the majority of the actual uses of the equations of chapter 3 are in the proofs which are to be found in appendix A. There we adopt a different style of assuming familiarity with the framework and giving *hints* for key steps.

[2] In particular MACE4.

over the set of arguments, and then as extended by Cayrol and Lagasquie-Schiex to a theory of bipolar argumentation, which adds a second relation of *support* [2, 26, 27] (or in our interpretation *explicit trust*) over the set of arguments. For each style of argumentation we show how to translate the argumentation scheme into a Boolean Network.

We then return to the theme of gaining initial trust and argue that the Cayrol and Lagasquie-Schiex conditions for bipolar networks are too strong for our application because they require an unreasonable degree of foresight of individuals making up the social matrix. To avoid this problem we provide a weaker set of conditions for bipolar networks that are aligned with a more limited degree of foresight. The resulting system provides a notion of *social connection* or *logical reputation* that aligns with our informal discussion above.

As a further development of this theory of initial trust, an aspect of the theory of privacy is developed as an example. The significance of privacy in modern, everyday, life is amply discussed by Wolfgang Sofsky in his book 'Privacy: A Manifesto" [108]. The technical view of privacy adopted for this example roughly follows the line taken in Helen Nissenbaum's "Privacy in Context" [78] and Daniel J. Solove's "Understanding Privacy" [109]. That is, privacy is contextual and governed by social norms that depend upon the context. The particular aspect of privacy chosen is that of making the judgement to disclose information. We will disclose private information to an individual only if there is a *need* to do so; only if we judge it *safe* to do so, in that the individual will handle the information appropriately according to the social norms for handling such information; and only if we believe the information will only be used for the *purpose* for which it is disclosed. Moreover part of the safety requirement is that the individual to whom we disclose the private information will apply the same rules in disclosing that information to third parties. We may assess the trustworthiness for disclosure by the Social Trust process discussed above. But we may also assess, given transparent access to the required information, how our private information could possibly flow from the individual to whom it is disclosed, to other members of the clubs to which that individual belongs. Since we assume the individual will follow the appropriate social norms in handling the information, and will only pass information on to those that are believed to follow the same social norms as determined by his club membership (i.e. the club distrust and explicit trust relations reflect the norm following behaviour), our assessment of appropriate privacy maintenance, as disclosers of the information, becomes whether, under these social norms, there is any possibility of an individual that we distrust becoming a recipient of the information. By changing terminology and labelling individuals who are *not distrusted*, as individuals that are *implicitly trusted*, the condition becomes simply that we implicitly trust everyone who may possibly handle the information.

5.2. Basic Abstract Argumentation Theory

Dung [34] introduces the idea that the important thing about arguments is how a collection of arguments work together to define what might be called a *position*. A

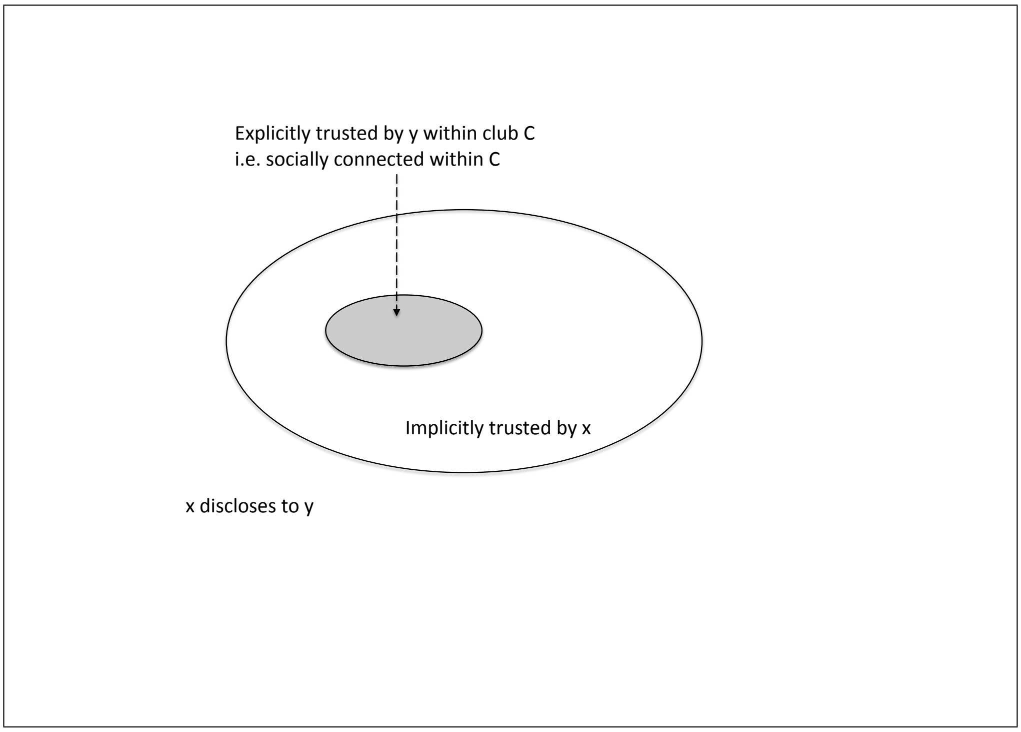
Explicitly trusted by y within club C
i.e. socially connected within C
Implicitly trusted by x
x discloses to y

Figure 5.2.: Privacy

position is a set of arguments that do not attack one another and defend themselves against attacks by other arguments, in the sense that if any argument attacks an argument which is part of the position, then some argument of the position counter attacks the attacker.

An argumentation system is a collection of arguments and a binary relation of attack between arguments. A semantics for an argumentation system is a prescription for how to resolve attacks between arguments to pick out viable *positions*. The meaning of an argumentation system, under a given semantics, is the set of all viable positions that one might adopt.

A Dungian argumentation system is formally modelled as a pair $(\mathbf{Args}, \mathbf{att})$ where **Args** is a set of arguments and **att** is a binary attack relation over **Args** such that $x\,\mathbf{att}\,y$ is read as x attacks y. Positions are formally modelled by the notion of an *Admissible set* which is defined as the conjunction of two other properties, that of being a *Conflict Free set* and that of being an *Acceptable set*. To define these notions it is convenient to define a derived relation of a set attacking an argument, written $S\,\mathbf{Att}\,x$, for a set S attacking an argument x, with the meaning that some member y of S attacks x (i.e. $\exists y \in S.\, y\,\mathbf{att}\,x$). The properties may then be stated as:

- *Conflict free*: A set of arguments $S \subseteq \mathbf{Args}$ is *conflict free* iff there is no pair of arguments $x \in S$ and $y \in S$ such that x **att** y.

- *Acceptable*: An argument $x \in \mathbf{Args}$ is *acceptable with respect to S* iff for every argument $y \in \mathbf{Args}$ if $y\,\mathbf{att}\,x$ then $S\,\mathbf{Att}\,y$. Following [12] we will also say that S *defends* x when x is acceptable with respect to S. Define

$D(X) = \{y \mid X \,\text{defends}\, y\}$.

- *Admissible*: A set $S \subseteq$ **Args** is *admissible* iff S is *conflict free* and each argument in S is *acceptable with respect to* S.

These conditions are illustrated in the following diagram:

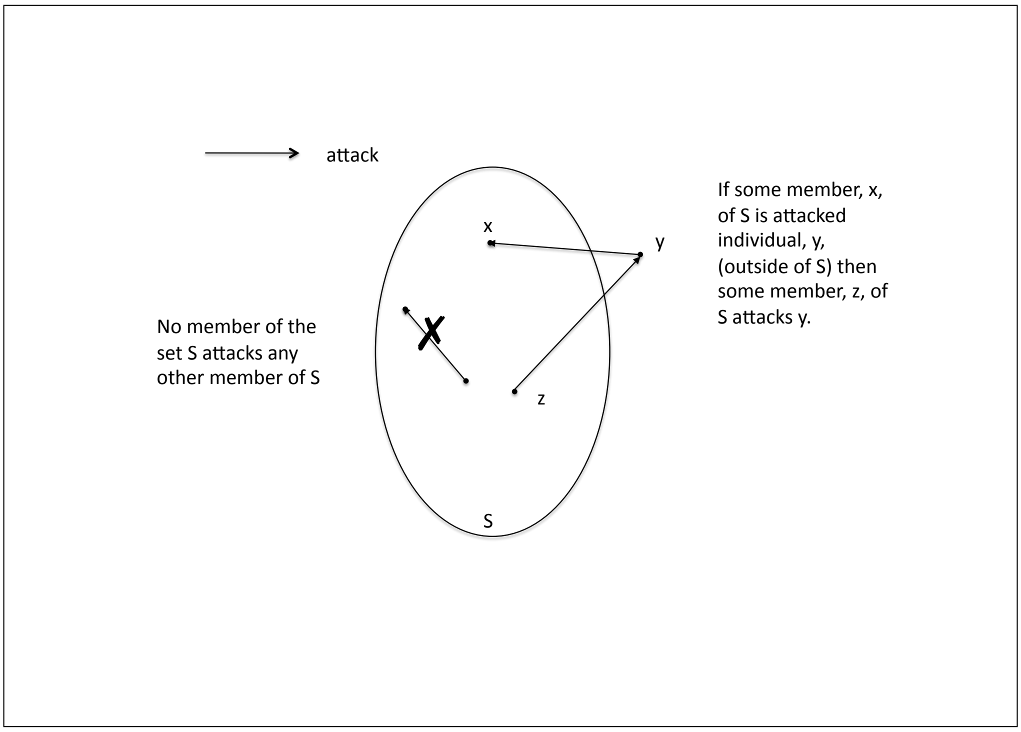
attack
x
y
z
S
If some member, x, of S is attacked individual, y, (outside of S) then some member, z, of S attacks y.
No member of the set S attacks any other member of S

Figure 5.3.: Conflict Free and Acceptable

Given a semantics, the positions generated by a given semantics are called the extensions of the argumentation system under the semantics.

In addition to Dung's original admissibility semantics, other semantics have been defined by adding additional constraints on what constitutes an extension.

These include:

- *complete extensions*: an admissible set, S, is a complete extension if it includes all arguments it defends i.e. $D(S) \subseteq S$.
- *grounded extensions*: an admissible set, S, is a ground extension if it is a minimal complete extension i.e. S is a complete extension and no complete extension is a proper subset of S.
- *preferred extensions*: an admissible set, S, is a preferred extension if it is a maximal complete extension i.e. S is a complete extension and no complete extension is a proper superset of S.
- *semi-stable extensions*: an admissible set, S, is a semi-stable extension if it is a preferred extension such that $S \cup \mathbf{att}(S)$ is maximal i.e. there is no preferred

extension, say T, such that $(S \cup \mathbf{att}(S)) \subset (T \cup \mathbf{att}(T))$ (let us call this ordering the *coverage ordering* for future reference).

- *stable extensions*: an admissible set, S, is a stable extension if it is a semi-stable extension such that $S \cup \mathbf{att}(S) = \mathbf{Args}$.

Formal accounts of the semantics of Dungian argumentation systems may be found in the book by Philippe Besnard and Anthony Hunter "Elements of Argumentation" [13], or in the paper "Characterization of Semantics for Argument Systems" by Besnard and Sylvie Doutre [12]. Additional insight may be gained by consulting Dung's original papers [34], the subsequent joint paper [17], and Leila Amgoud and Claudette Cayrol's "On the Acceptability of Arguments in Preference-based Argumentation" [1].

5.3. Abstract Argumentation Theory and Galois Connections

It is quite apparent, even in the limited context of Dung's framework, that there are many alternative semantics for argumentation systems. And the possible variety increases as we move to bipolar systems and beyond. Potentially this variety complicates the study of both theoretical and practical aspects of argumentation. Theoretically it gives rise to questions of how the various semantics correspond to one another. Practically it gives rise to the need for a variety of evaluation algorithms to find solution sets (sets of positions) for particular argumentation systems against given semantics. To help address these issues a systematic approach to the semantics is presented that allows many semantics to be expressed in a standard form and an evaluation strategy is presented that allows us to evaluate argumentation systems against a semantics written in this standard form.

The ideas are first illustrated for Dung-style systems and then defined for a class of generalised argumentation systems. Later these generalised systems are used to express the semantics of standard bipolar argumentation systems and of trust systems, a variation on the theme of bipolar systems.

Firstly we rewrite the various semantics for Dung's systems into relational operator style using the axiality $(R_{\exists}, R^{\forall})^{[3.89]}$. Following [12], conflict freeness and acceptability may be phrased as:

- *Conflict Free*: $R_{\exists}S \subseteq \overline{S}$

- *Acceptable*: $\breve{R}_{\exists}S \subseteq R_{\exists}S$

and, $D(S)$, the set of items defended by a set S may be phrased as:

- $D(S) \mathrel{\widehat{=}} \{y \mid \breve{R}_{\exists}\{y\} \subseteq R_{\exists}S\}$

Using this Galois Connection we phrase equivalents of acceptability and completeness (proof for this chapter are supplied in appendix A):

PROPOSITION 1

- *Acceptability - defined as $\breve{R}_\exists S \subseteq R_\exists S$ is equivalently expressed as $S \subseteq \breve{R}^\forall R_\exists S$.*
- *Completness - defined as $D(S) \subseteq S$, is equivalently expressed as $\breve{R}^\forall R_\exists S \subseteq S$.*

Hence S is acceptable and complete if, and only if, $\breve{R}^\forall R_\exists S = S$.

PROPOSITION 2 *S is conflict free is equivalent to $\breve{R}_\exists S \subseteq \overline{S}$*

We next turn to stable extensions. A stable extension fulfils the condition $S \cup R_\exists S = \textbf{Args}$. We show this is equivalent to $\overline{S} \subseteq R_\exists S$.

PROPOSITION 3 *$S \cup R_\exists S = \textbf{Args}$ is equivalent to $\overline{S} \subseteq R_\exists S$.*

Moreover, if S is admissible and satisfies $S \cup R_\exists S = \textbf{Args}$, then it is preferred and semi-stable:

PROPOSITION 4 *If S is admissible and $S \cup R_\exists S = \textbf{Args}$ then it is preferred and semi-stable.*

Hence if S is admissible and $\overline{S} \subseteq R_\exists S$ then S is a stable extension.

Thus we see a pleasing symmetry in the conditions used to define extensions:

conflict free	$R_\exists S \subseteq \overline{S}$
acceptable	$S \subseteq \breve{R}^\forall R_\exists S$
complete	$\breve{R}^\forall R_\exists S \subseteq S$
stable	$\overline{S} \subseteq R_\exists S$

That is, an admissible set of an argumentation systems is any set $S \subseteq \textbf{Args}$ satisfying conflict free and acceptable i.e. the set of conditions $\{R_\exists S \subseteq \overline{S}, S \subseteq \breve{R}^\forall R_\exists S\}$. Similarly S satisfies a complete semantics if it satisfies conflict free, acceptable and complete, i.e. the set of conditions $\{R_\exists S \subseteq \overline{S}, S \subseteq \breve{R}^\forall R_\exists S, \breve{R}^\forall R_\exists S \subseteq S\}$. And finally, S satisfies stable semantics if it satisfies conflict free, acceptable, complete and stable i.e. $\{R_\exists S \subseteq \overline{S}, S \subseteq \breve{R}^\forall R_\exists S, \breve{R}^\forall R_\exists S \subseteq S, \overline{S} \subseteq R_\exists S\}$.

Some simplification is possible in this last case since:

PROPOSITION 5 *If S is conflict free and stable then it is acceptable and complete.*

That is, S satisfies stable semantics if, and only if, S satisfies conflict free and stable i.e. it satisfies the conditions $\{R_\exists S \subseteq \overline{S}, \overline{S} \subseteq R_\exists S\}$.

The *Ground*, *Preferred* and *Semi-Stable* semantics are derived from these four conditions by the use of minimisation and maximisation.

Given an argumentation system (A, R), the set of complete positions is defined by:

- Complete$(A) \mathrel{\widehat{=}} \{S \subseteq A \mid S = \breve{R}^\forall R_\exists S \wedge R_\exists S \subseteq \overline{S}\}$

then the sets of ground and preferred positions are defined by:

- Ground$(A) \mathrel{\widehat{=}} \{S \in \text{Complete}(A) \mid \neg\, \exists\, S' \in \text{Complete}(A).S' \subset S\}$

- Preferred$(A) \mathrel{\widehat{=}} \{S \in \text{Complete}(A) \mid \neg\, \exists\, S' \in \text{Complete}(A).S \subset S'\}$

and defining the coverage ordering, $\Subset$ by:

- $A \Subset B \mathrel{\widehat{=}} A \cup R_{\exists}A \subset B \cup R_{\exists}B$

the *Semi-Stable* sets are given by:

- SemiStable$(A) \mathrel{\widehat{=}} \{S \in \text{Preferred}(A) \mid \neg\, \exists\, S' \in \text{Preferred}(A).S \Subset S'\}$

Thus, in the case of Dung's systems, many forms of semantics can be given by specifying a set of constraints on solution sets and performing minimisation or maximisation with respect to an ordering. The semantics so expressed makes plain much of the correspondence between the various semantics.

We next consider how to evaluate such systems.

5.4. Boolean Network Approach to Argumentation Theory

In this section a notion of generalised argumentation system is defined, along with a translation, from an argumentation system into a system of propositional variables and propositional axioms, which is guided by the semantics. The propositional axioms may be thought of as defining a Boolean Network in which the propositional variables represent the "state" of nodes and the propositional axioms represent how the nodes are wired together via boolean functions. In particular this network incorporates feedback i.e. it may contain cyclic linkages. A model of a boolean network is a propositional model that satisfies the axioms. The translation from the argumentation is such that models of the Boolean Network correspond to positions of the argumentation system.

To briefly motivate the approach, consider the simple Dungian argumentation system depicted in figure 5.4(i) and take admissibility semantics.

In the translation, the arguments will be replaced by Boolean variables a, b and c. The truth value of the boolean variables represents whether or not the corresponding argument is part of a given position. So a true, and both b and c false, represents the position $\{A\}$, and a and c true and b false represents the position $\{A, C\}$. Conflict between arguments corresponds to inhibition between boolean variables, where by inhibition we mean x inhibits y if x being true implies y is false.

The example argument structure may be represented as a network of boolean variables connected by inhibition (as shown in figure 5.4(ii)).

In addition to inhibition between variables, Dungian argumentation adds the acceptability condition which requires a position to defend itself. Translating the acceptability condition to a constraint on truth values results in the defence condition that if a boolean variable x is true then each potential inhibitor y_i of x must be inhibited by some inhibitor z^i_j of y_i i.e. $\bigwedge_i \bigvee_j z^i_j$ must be true.

Applied to the above this requires that

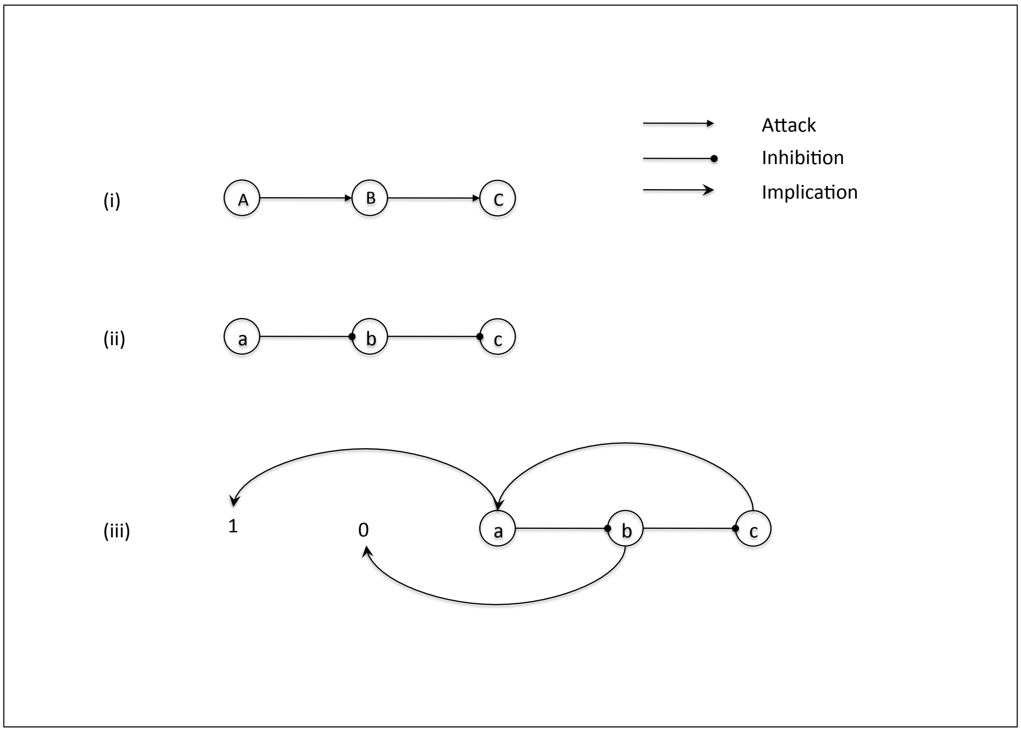

Figure 5.4.: Argumentation and Boolean Networks

- if c is true then a is true,
- if b is true then $\bigwedge_i \bigvee_j \emptyset$, which is false, so b is false,
- if a is true then $\bigwedge_i \emptyset$, which is true

These additional constraints are depicted in figure 5.4(iii).

An assignment of truth values to the network is valid provided that it meets the inhibition conditions and the Dungian defence conditions of the network. Valid assignments correspond to admissible positions of the argumentation system.

It is perhaps worth noting that the backward constraint $c \Rightarrow a$ can always be converted to a forward constraint using the NOR connective (x NOR y meaning neither of x or y is true) to derive an inhibition network. This is illustrated using a slightly more complicated argumentation system in which, B, of the previous example, is given two attackers A_1 and A_2. This is illustrated in figure 5.5 starting with the translation into the initial system of boolean variables 5.5(i) and then continuing with the addition of the defence constraints 5.5(ii). Figure 5.5(iii) illustrates the network with NOR's and figure 5.5(iv) illustrates the simplification, in which it is seen that neither a_1 nor a_2 are ever inhibited and b is always inhibited. This means that a_1 and a_2 may freely take on values true or false, b must always take the value false and c may take the value true only if either, or both, of a_1 and a_2 take the value true.

The simple translation process illustrated here generalises over semantics defined by relational operator inequalities (subset inclusions). This is true not only

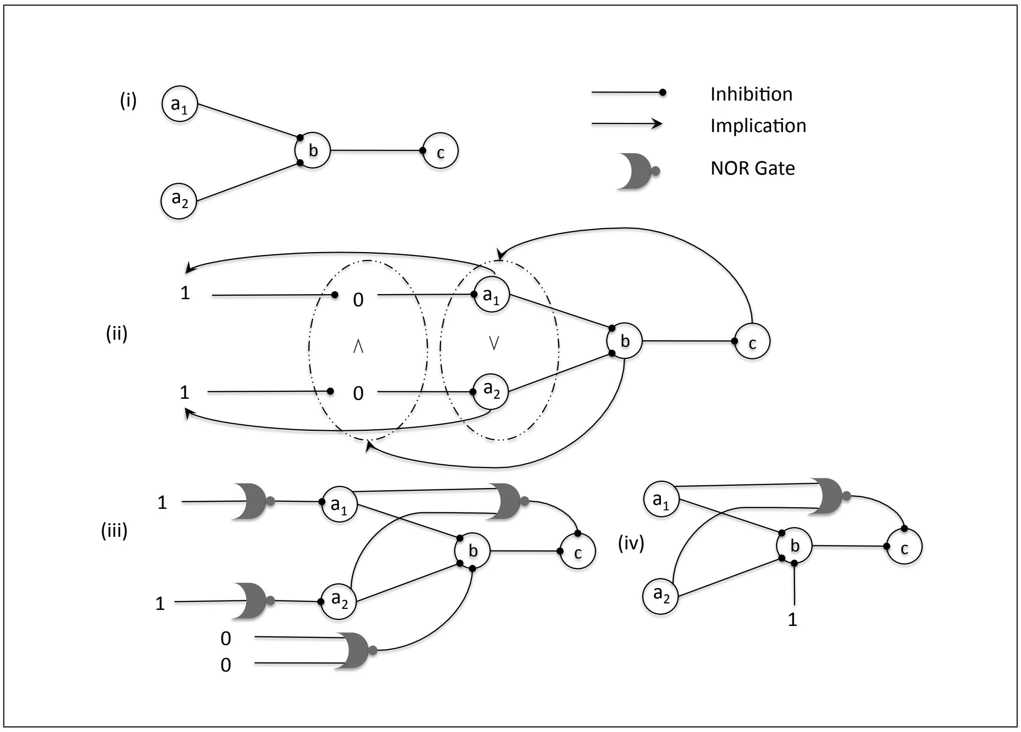

Figure 5.5.: Boolean Networks and Forward Constraints

for Dungian argumentation systems but for a wider class of systems, called here, generalised argumentation systems.

After defining the generalised argumentation system we give a compilation of the various semantics for Dung's systems and illustrate the process of finding solution sets by taking some example Dung style argumentation systems under the various semantics.

The distinction between semantics which only require constraint satisfaction versus those which require a maximisation (minimisation) step occurs again in finding models for the Boolean Networks. If a semantics does not require maximisation or minimisation we will call it a local semantics, otherwise we will call it global. With local semantics models are defined without reference to other models. In the semantics for Dung's systems, these semantics are also local in another sense, in that the acceptability of an argument in a set of arguments can be calculated from *nearby* arguments.

The basic tool we use is a recursive translation of the argumentation conditions into boolean expressions.

A generalised argumentation system is a tuple $(\mathbf{Args}, J_1, \ldots, J_n)$ where **Args** is some set of arguments and each J_i is a binary relation. A semantics for an argumentation system is a collection of set inclusions conditions $C = \{C_i \mid i \in I\}$, where each C_i is of the form $A = B$ or $A \subseteq B$ where the sets A and B are defined by free variables, set operators and relational operators as in the examples of the semantics of Dung' systems. The complete form is given in figure 5.6.

A solution is a collection of subsets of **Args**, S_x, one for each set variable x

A set is one of:

- The empty set $\emptyset$.
- The set of all arguments **Args**.
- Some set variable, e.g. x.
- The union of sets e.g. $X \cup Y$.
- The intersection of sets e.g. $X \cap Y$.
- The complement of a set e.g. $\overline{X}$.
- The image of a set under the operator $R_{\exists}$ for some binary relation R.
- The image of a set under the operator $R^{\forall}$ for some binary relation R.

And a binary relation is one of:

- a binary relation J_i from the argumentation system.
- union, e.g. $R_1 \cup R_2$.
- intersection, e.g. $R_1 \cap R_2$.
- complementation e.g. $\overline{R}$.
- converse e.g. $\breve{R}$.

Figure 5.6.: Sets and Relations

mentioned in C, that simultaneously meet all the conditions in C. We will refer to the solutions as the meanings of $(\mathbf{Args}, J_1, \ldots, J_n)$ assigned by C.

Given a solution S, a full model for the argumentation system is a set of characteristic functions for the solution i.e. a set of functions $f_x : \mathbf{Args} \rightarrow \{0, 1\}$, one for each variable x mentioned in the semantics, such that, $\forall\, y \in \mathbf{Args}\,.\; (f_x(y) = 1) \equiv y \in S_x$.

Usually we are interested in only a single free variable that represents the positions of the argumentation system under the semantics, other variables being auxiliary, construction, variables. By convention we will reserve the set variable $\mathcal{X}$ for the set of positions in the semantics. This means that each solution set for this variable corresponds to a possible *position* of the argumentation system. A model for an argumentation system will be the characteristic function for $\mathcal{X}$, i.e. a function mapping each argument to the Booleans 1 (i.e. **true**) and 0 (i.e. **false**), with 1 indicating that the argument is part of the position and 0 indicating it is no part of the position.

Given an argumentation system $(\mathbf{Args}, J_1, \ldots, J_n)$ we associate a propositional system (P, Ax), where P is a set of propositional variables containing a variable x_a for each free variable x in C and each argument $a \in \mathbf{Args}$, and Ax is a set of axioms derived from the particular semantics C under consideration. The essence of the correspondence is that each propositional variable S_a, of the propositional system, corresponds to the assertion $a \in S$ of the argumentation system. Each semantic condition is ultimately expanded out into propositional connections between these elementary membership propositions. The recursive transformation of the conditions, C, of the argumentation system's semantics into the set of axioms, Ax, of the propositional system is given in figure 5.7.

If S is a set variable of C then $\phi_S : \mathbf{Args} \rightarrow \mathit{PropVar}$ is the injection between **Args** and the propositional variables, so that, $\phi_S(a) = S_a$ means that S_a is the propositional variable corresponding to argument a being in the set S.

PROPOSITION 6 *A (propositional) model m of C^T defines a full model of C by $f_S \mathrel{\widehat{=}} m \circ \phi_S$ for each set variable S mentioned in C.*

This follows directly from the construction by use of the equivalences.

Non-local semantics requires minimising or maximising of solution sets with respect to one or more relations. The process is carried out sequentially i.e. first optimising the solution sets with respect to one relation and then optimising the remaining solution sets with respect to the next. To this end a non-local semantics is given by a set of local conditions and a sequence of optimisation conditions read as maximising conditions[3]. Optimisation is performed over the solution set $\mathcal{X}$. Thus, for example, preferential semantics is defined by the tuple $(C, \subseteq)$ with C, the conditions for conflict free and complete solutions, maximised for set inclusion. And semi-stable theories being given by $(C, \subseteq, \prec)$ where $\prec$ is the stability ordering[4], which is used to select maximum sets from the preferred extensions.

[3] Minimisation is performed by using the complementary ordering.

[4] i.e. $f \prec g \equiv \{x \mid f \circ \phi_{\mathcal{X}}(x) = 1\} \subset \{x \mid g \circ \phi_{\mathcal{X}}(x) = 1\}$.

$$(X = Y)^T \mathrel{\widehat{=}} \bigwedge_x ((x \in X)^T \equiv (x \in Y)^T) \tag{5.1}$$

$$(X \subseteq Y)^T \mathrel{\widehat{=}} \bigwedge_x ((x \in X)^T \Rightarrow (x \in Y)^T) \tag{5.2}$$

$$(x \in \emptyset)^T \mathrel{\widehat{=}} \mathbf{false} \tag{5.3}$$

$$(x \in \mathbf{Args})^T \mathrel{\widehat{=}} \mathbf{true} \tag{5.4}$$

$$(x \in X \cup Y)^T \mathrel{\widehat{=}} (x \in X)^T \vee (x \in Y)^T \tag{5.5}$$

$$(x \in X \cap Y)^T \mathrel{\widehat{=}} (x \in X)^T \wedge (x \in Y)^T \tag{5.6}$$

$$(x \in \overline{X})^T \mathrel{\widehat{=}} \neg\, (x \in X)^T \tag{5.7}$$

$$(x \in J_{\exists} X)^T \mathrel{\widehat{=}} \bigvee_{((y,x) \in J)^T} (y \in X)^T \tag{5.8}$$

$$(x \in J^{\forall} X)^T \mathrel{\widehat{=}} \bigwedge_{((y,x) \in \breve{J})^T} (y \in X)^T \tag{5.9}$$

$$(x \in S)^T \mathrel{\widehat{=}} S_x \text{ for } S \text{ a set variable} \tag{5.10}$$

where $((y,x) \in J)^T$, for J a compound relation is defined by:

$$((y,x) \in J_1 \cup J_2)^T \mathrel{\widehat{=}} ((y,x) \in J_1)^T \vee ((y,x) \in J_2)^T \tag{5.11}$$

$$((y,x) \in J_1 \cap J_2)^T \mathrel{\widehat{=}} ((y,x) \in J_1)^T \wedge ((y,x) \in J_2)^T \tag{5.12}$$

$$((y,x) \in \bar{J})^T \mathrel{\widehat{=}} (y,x) \notin J \tag{5.13}$$

$$((y,x) \in \breve{J})^T \mathrel{\widehat{=}} (x,y) \in J \tag{5.14}$$

and define

$$\{C_i \mid i \in I\}^T \mathrel{\widehat{=}} \{C_i^T \mid i \in I\} \tag{5.15}$$

Figure 5.7.: Translation

These maximisations are performed by selecting the model that maximises the required set of assignments in the boolean model.

5.5. Modelling Dungian Systems

We now look at the translation of the semantics for Dung's systems into their propositional system equivalents.

Dungian systems are defined as argumentation systems with a single relation of attack between arguments i.e. $(\mathbf{Args}, \mathbf{att})$ with the basic semantics being admissibility semantics defined by:

$$C_{\text{admissable}} = \{\mathbf{att}_{\exists}\,\mathcal{X} \subseteq \overline{\mathcal{X}}, \breve{\mathbf{att}}_{\exists}\mathcal{X} \subseteq \mathbf{att}_{\exists}\,\mathcal{X}\} \tag{5.16}$$

which corresponds to:

$$\bigwedge_{x \in \mathbf{Args}} (\bigvee_{(y,x) \in A} \mathcal{X}_y \Rightarrow \neg\, \mathcal{X}_x) \tag{5.17}$$

$$\bigwedge_{x \in \mathbf{Args}} (\bigvee_{(x,y) \in A} \mathcal{X}_y \Rightarrow \bigvee_{(z,x) \in A} \mathcal{X}_z) \tag{5.18}$$

or, since we can write acceptability as $\mathcal{X} \subseteq \breve{\mathbf{att}}^{\forall}\,\mathbf{att}_{\exists}\,\mathcal{X}$, we can translate the semantics equivalently as:

$$\bigwedge_{x \in \mathbf{Args}} (\bigvee_{(y,x) \in A} \mathcal{X}_y \Rightarrow \neg\, \mathcal{X}_x) \tag{5.19}$$

$$\bigwedge_{x \in \mathbf{Args}} (\mathcal{X}_x \Rightarrow \bigwedge_{(y,x) \in A} \bigvee_{(z,y) \in A} \mathcal{X}_z) \tag{5.20}$$

(which, of course, we may also derive via propositional reasoning from the former by "moving the disjunction to the other side of the implication").

As we saw above, complete extensions are obtained by adding the dual condition to acceptability which turns the acceptability condition into an equivalence i.e.

$$C_{\text{complete}} = \{\mathbf{att}_{\exists}\,\mathcal{X} \subseteq \overline{\mathcal{X}}, \mathcal{X} = \breve{\mathbf{att}}^{\forall}\,\mathbf{att}_{\exists}\,\mathcal{X}\} \tag{5.21}$$

yielding the Boolean conditions:

$$\bigwedge_{x \in \mathbf{Args}} (\bigvee_{(y,x) \in A} \mathcal{X}_y \Rightarrow \neg\, \mathcal{X}_x) \tag{5.22}$$

$$\bigwedge_{x \in \mathbf{Args}} (\mathcal{X}_x \equiv \bigwedge_{(y,x) \in A} \bigvee_{(z,y) \in A} \mathcal{X}_z) \tag{5.23}$$

and stable extensions are obtained by adding the dual of the conflict freeness condition turning this into an equivalence:

$$C_{\text{stable}} = \{\mathbf{att}_{\exists}\, \mathcal{X} = \overline{\mathcal{X}}, \mathcal{X} = \breve{\mathbf{att}}^{\forall}\, \mathbf{att}_{\exists}\, \mathcal{X}\} \tag{5.24}$$

yielding:

$$\bigwedge_{x \in \mathbf{Args}} \Big(\bigvee_{(y,x) \in A} \mathcal{X}_y \equiv \neg\, \mathcal{X}_x \Big) \tag{5.25}$$

$$\bigwedge_{x \in \mathbf{Args}} \Big(\mathcal{X}_x \equiv \bigwedge_{(y,x) \in A} \bigvee_{(z,y) \in A} \mathcal{X}_z \Big) \tag{5.26}$$

Moreover since above it was shown that conflict free and stable sets are acceptable and complete then the only required condition is:

$$\bigwedge_{x \in \mathbf{Args}} \Big(\bigvee_{(y,x) \in A} \mathcal{X}_y \equiv \neg\, \mathcal{X}_x \Big) \tag{5.27}$$

Using the extended notion of a semantics, the ground extensions and preferred extensions are defined by maximisation of complete semantics with respect to $\overline{\mathcal{X}}$ and $\mathcal{X}$ respectively i.e. $(C_{\text{complete}}, \{\mathcal{X}\})$ and $(C_{\text{complete}}, \{\mathcal{X}\})$. The semi-stable semantics is defined by maximising with respect to the set $\mathbf{att}_{\exists}\, \mathcal{X} \cup \mathcal{X}$ i.e. $(C_{\text{complete}}, \{\mathbf{att}_{\exists}\, \mathcal{X} \cup \mathcal{X}\})$.

5.6. Computing Some Dungian Examples

We will describe models by giving the set of propositional letters assigned true by the assignment i.e. for a set of letters $\{A, B, C, D\}$, the set $\{A, C\}$ represents the model that assigns A and C **true**, and B and D **false**.

If we take a classic set of argumentation examples, such as shown in figure 5.8 (i) - (iv), the extended implication network codings are:

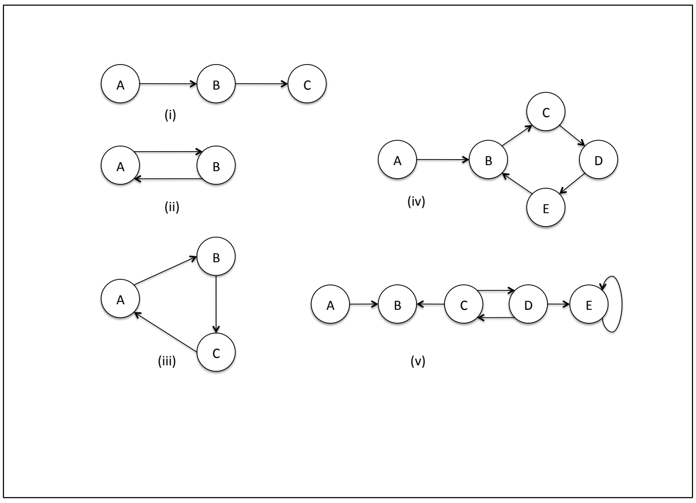

Figure 5.8.: Example Argumentation Structures

sub-figure	Conflict free	Acceptable	Models
(i)	$\textbf{false} \Rightarrow \neg A$ $A \Rightarrow \neg B$ $B \Rightarrow \neg C$	$C \Rightarrow A$ $B \Rightarrow \textbf{false}$	$\emptyset$, $\{A\}$, $\{A, C\}$
(ii)	$A \Rightarrow \neg B$ $B \Rightarrow \neg A$	$B \Rightarrow B$ $A \Rightarrow A$	$\emptyset$, $\{A\}$, $\{B\}$
(iii)	$A \Rightarrow \neg B$ $B \Rightarrow \neg C$ $C \Rightarrow \neg A$	$B \Rightarrow C$ $C \Rightarrow A$ $A \Rightarrow B$	$\emptyset$
(iv)	$\textbf{false} \Rightarrow \neg A$ $A \Rightarrow \neg B$ $B \Rightarrow \neg C$ $C \Rightarrow \neg D$ $D \Rightarrow \neg E$ $E \Rightarrow \neg B$	$C \Rightarrow (A \vee E)$ $B \Rightarrow \textbf{false}$ $B \Rightarrow D$ $E \Rightarrow C$ $D \Rightarrow E$	$\emptyset$, $\{A\}$, $\{A, C\}$, $\{C, E\}$, $\{A, C, E\}$
(v)	$\textbf{false} \Rightarrow \neg A$ $A \Rightarrow \neg B$ $C \Rightarrow \neg B$ $C \Rightarrow \neg D$ $D \Rightarrow \neg C$ $D \Rightarrow \neg E$ $E \Rightarrow \neg E$	$B \Rightarrow \textbf{false}$ $B \Rightarrow D$ $C \Rightarrow C$ $D \Rightarrow D$ $E \Rightarrow C$ $E \Rightarrow E$	$\emptyset$, $\{A\}$, $\{C\}$, $\{D\}$, $\{A, C\}$, $\{A, D\}$

sub-figure	Complete	Models	Stable	Models
(i)	**true** $\Rightarrow A$ $A \Rightarrow C$	$\{A, C\}$	$\neg\, A \Rightarrow$ **false** $\neg\, A \Rightarrow B$ $\neg\, B \Rightarrow C$	{A, C}
(ii)	$A \Rightarrow A$ $B \Rightarrow B$	$\emptyset, \{A\}, \{B\}$	$\neg\, A \Rightarrow B$ $\neg\, B \Rightarrow A$	$\{A\}, \{B\}$
(iii)	$B \Rightarrow A$ $C \Rightarrow B$ $A \Rightarrow C$	$\emptyset$	$\neg\, A \Rightarrow B$ $\neg\, B \Rightarrow C$ $\neg\, C \Rightarrow A$	-no model-
(iv)	**true** $\Rightarrow A$ $A \Rightarrow C$ $B \Rightarrow D$ $D \Rightarrow B$ $E \Rightarrow C$	$\{A, C, E\}$	$\neg\, A \Rightarrow$ **false** $\neg\, (A \vee E) \Rightarrow B$ $\neg\, B \Rightarrow C$ $\neg\, C \Rightarrow D$ $\neg\, D \Rightarrow E$	$\{A, C, E\}$
(v)	**true** $\Rightarrow A$ (**false** $\wedge\, D) \Rightarrow B$ $C \Rightarrow C$ $D \Rightarrow D$ $(C \wedge D \wedge E) \Rightarrow E$	$\{A\}, \{A, C\},$ $\{A, D\}$	$\neg\, A \Rightarrow$ **false** $\neg\, (A \vee C) \Rightarrow B$ $\neg\, D \Rightarrow C$ $\neg\, C \Rightarrow D$ $\neg\, (D \vee E) \Rightarrow E$	$\{A, D\}$

5.7. A Note on the Generalised Argumentation Systems

Above, argumentation systems were generalised to allow multiple relations. Below this generalisation will be used for Bipolar argumentation systems. However, to illustrate some of the potential for such generalisation consider the following form of argumentation. An argument may attack another argument either *factually* or *morally*. An argument may both attack some arguments factually and other arguments morally. A collection of arguments is regarded as a plausible position if it defends itself against factual arguments by factual counter attacks and moral arguments by moral counter attacks. The formulation of this system is quite straightforward: a plausible position is required to be both factually and morally conflict free and that any argument which is factually attacked is factually defended and any argument which is morally attacked is morally defended. Let $(\mathbf{Args}, F, M)$ be the generalised argumentation system with semantics:

$$F_{\exists}\mathcal{X} \subseteq \overline{\mathcal{X}} \tag{5.28}$$

$$M_{\exists}\mathcal{X} \subseteq \overline{\mathcal{X}} \tag{5.29}$$

$$\mathcal{X} \subseteq \breve{F}^{\forall} F_{\exists}\mathcal{X} \tag{5.30}$$

$$\mathcal{X} \subseteq \breve{M}^{\forall} M_{\exists}\mathcal{X} \tag{5.31}$$

If, however, we wanted to allow a different notion of defence then we could, for example, allow factual attacks to be either morally or factually refuted but require moral attacks to be morally refuted. The resulting semantics is:

$$\begin{aligned} F_\exists\mathcal{X} &\subseteq \overline{\mathcal{X}} && (5.32)\\ M_\exists\mathcal{X} &\subseteq \overline{\mathcal{X}} && (5.33)\\ \mathcal{X} &\subseteq \breve{F}^\forall(F_\exists\mathcal{X} \cup M_\exists\mathcal{X}) && (5.34)\\ \mathcal{X} &\subseteq \breve{M}^\forall M_\exists\mathcal{X} && (5.35)\\ & && (5.36) \end{aligned}$$

5.8. Adding Priority between Arguments

A straightforward extension of Dung's argumentation system is to add a simple priority relation between arguments.

Priority between arguments expresses the notion that one argument is stronger than another. The idea is that if one argument attacks another then it is the stronger argument that wins. This may be expressed as a generalised argumentation system obtained by adding a second binary relation to a Dungian system to represent the relative strength between arguments i.e. $(\mathbf{Args}, \mathbf{att}, \succ)$. The semantics are essentially those of Dungian systems with the semantic attack relation being taken as the conjunction with the two given relations and so is equivalent to the Dungian system $(\mathbf{Args}, \mathbf{att} \cap \succ)$ (i.e. an attack succeeds if the attacker is stronger than the argument attacked).

In some systems it is possible to separate out the constraints and express them as constraints on the individual given relations. However, in this case, without further constraints on **att** and/or $\succ$, no further simplification is possible since we have $(\mathbf{att} \cap \succ)_\exists X \subseteq \mathbf{att}_\exists\, X \cap (\succ)_\exists X$.

5.9. Bipolar Argumentation Systems

Bipolar argumentation systems are generalised argumentation systems with two binary relations, representing attack and support between arguments of the system i.e. they are triples $(\mathbf{Args}, \mathbf{att}, \mathbf{Support})$. They were originally introduced by Cayrol and Lagasquie-Schiex as a tool for modelling real world argumentation processes involving both positive and negative preferences over arguments.

In a later section, 5.12, our goal is to use bipolar argumentation systems as a basis for a theory of trust and distrust between individuals. In line with this end we will develop the theory of bipolar argumentation using the terms *(explicit) distrust* for the attack relation and *(explicit) trust* for the support relation. This should not cause too much disruption when thinking about abstract argumentation and, indeed, one can phrase the attack relation between arguments A and B as, *A causes one to distrust B*, and the support relation as *A causes one to trust B*.

The general approach to bipolar argumentation is illustrated by a simple example. Consider the set of argument $X = \{\mathbf{A}, \mathbf{B}\}$ where **A** trusts some argument **Z** and **B** distrusts **Z**. One may regard X as *internally* consistent because **A** and **B** do not distrust each other and, although they do not *explicitly* trust each other we might say, because of the lack of *explicit* distrust, that they implicitly trust each other. However, if one considers **A** and **B** in relation to **Z**, one might say X is *externally* inconsistent in that members of X have differing views on the trustworthiness of individuals not in X. External inconsistency indicates a lack of agreement on trust and distrust relationships by the members of X.

There are many, different, possible notions of internal and external consistency. The differences in their definitions can be intuitively thought of as involving some idea of *range of effect*. For example, internal consistency could simply mean *Conflict Free* i.e. no direct contradictions. But it could also mean, for example, that no member of the set indirectly trusts some non-member (i.e. via a chain of trust connections between non-members) that distrusts a member of the set. The subject of bipolar argumentation is really a matter of defining suitable notions of internal and external consistency.

We start by developing Cayrol and Lagasquie-Schiex's bipolar argumentation theory using Galois Connections and then define a new system that is more fully aligned to use in discussing trust and distrust. We will refer to the Cayrol and Lagasquie-Schiex bipolar framework as the *standard* system and the new system we introduce as *trust* systems.

In formulae we will still use A for the distrust relation, to avoid confusion with $D(S)$, the set of arguments defended by S, and use T to denote the trust relation.

First recall that for a relation, R, R^* is the reflexive, transitive closure of R[3.48]. For standard systems we introduce the definitions:

- *trust closed*: A set is trust closed if, and only if, $T^*_\exists X \subseteq X$.

Recall that a set X is conflict free if, and only if, $A_\exists X \subseteq \overline{X}$. Bipolar systems have the potential for additional consistency constraints between a set X and the trust closure of X.

- *internally trust consistent*: A set is internally trust consistent if, and only if, $(A \circ T^*)_\exists X \subseteq \overline{X}$.

That is, nothing in the trust closure of X attacks X.

- *externally trust consistent*: A set is externally trust consistent if, and only if, $(A \circ T^*)_\exists X \subseteq \overline{T^*_\exists X}$.

That is, the trust closure of X does not attack itself (and so also does not attack X, which follows from the next proposition).

These conditions are illustrated in figure 5.9. The following propositions formalise a number of 'obvious' diagrammatic conclusions.

PROPOSITION 7 $X \subseteq T^*_\exists X$

We now give a series of propositions which characterise the relationship between simple consistency (conflict freeness), internal consistency and external consistency.

PROPOSITION 8 *X is internally trust consistent is equivalent to: the attackers of X are in the complement of the trust closure of X i.e.* $\breve{A}_\exists X \subseteq \overline{T^*_\exists X}$*.*

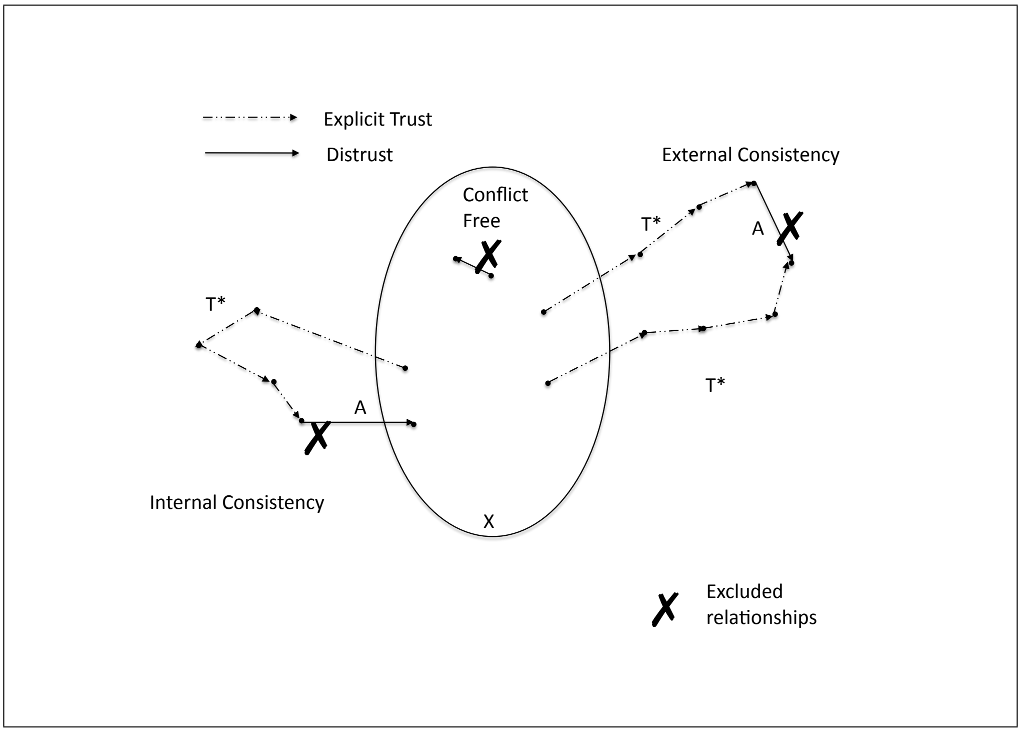

Figure 5.9.: Bipolar Conditions

PROPOSITION 9 *X is externally trust consistent is equivalent to: the attackers of the trust closure of X are in the complement of the trust closure i.e.* $\breve{A}_{\exists}(T^{*}_{\exists}X) \subseteq \overline{T^{*}_{\exists}X}$.

PROPOSITION 10 *A set X is externally trust consistent if its trust closure is conflict free, i.e.* $A_{\exists}(T^{*}_{\exists}X) \subseteq \overline{T^{*}_{\exists}X}$.

PROPOSITION 11 *If a set is externally trust consistent then it is internally trust consistent.*

PROPOSITION 12 *If a set is internally trust consistent then it is conflict free.*

PROPOSITION 13 *If X is trust closed and internally trust consistent then it is externally trust consistent.*

PROPOSITION 14 *X is externally trust consistent is equivalent to:*

A: $X \subseteq T^{*\forall}\breve{A}^{\forall} \circ \breve{T}^{*\forall}\overline{X}$ *and hence to*

B: $\breve{T}^{*}{}_{\exists}\breve{A}_{\exists}T^{*}{}_{\exists}X \subseteq \overline{X}$ *and*

C: $T^{*}{}_{\exists}A_{\exists}\breve{T}^{*}{}_{\exists}X \subseteq \overline{X}$

The semantics of the standard system is given by modifying the definition of admissibility from Dung's system by taking different consistency conditions and either basing acceptability on Dung's definition, i.e. $\breve{A}_\exists X \subseteq A_\exists X$, or basing acceptability on the trust closure of a set i.e. $\breve{A}_\exists(T^*_\exists X) \subseteq A_\exists(T^*_\exists X)$.

These semantics differ in how they treat arguments that support arguments that are ultimately rejected. Internal consistency simply requires that the set of trusted arguments have no conflicts between themselves, either directly or through some chain of trusting other arguments. Whereas external consistency means that any argument that would come into conflict by following chains of trust is regarded as itself being in conflict i.e. backers of ultimately untrusted arguments become untrusted.

We next define various notions of admissibility and show how they relate to one another.

- *Basic Dungian admissibility*: Dungian Conflict Free and Dungian Acceptable i.e. $\textbf{admissible}_D\, X \equiv A_\exists X \subseteq \overline{X} \wedge \breve{A}_\exists X \subseteq A_\exists X$.

- *Internal admissibility*: Internally Consistent and Dungian Acceptable i.e. $\textbf{admissible}_I\, X \equiv A_\exists T^*_\exists X \subseteq \overline{X} \wedge \breve{A}_\exists X \subseteq A_\exists X$.

- *External admissibility*: Externally Consistent and Dungian Acceptable i.e. $\textbf{admissible}_E\, X \equiv A_\exists T^*_\exists X \subseteq \overline{T^*_\exists X} \wedge \breve{A}_\exists X \subseteq A_\exists X$.

- *Trust Closed admissibility*: Trust Closed and Dungian Admissable i.e. $\textbf{admissible}_T\, X \equiv T^*_\exists X \subseteq X \wedge \textbf{admissible}_D\, X$.

- *Trust Extended admissibility*: Dungian Admissibility of the Trust Closure i.e. $\textbf{admissible}_X\, X \equiv A_\exists T^*_\exists X \subseteq \overline{T^*_\exists X} \wedge \breve{A}_\exists T^*_\exists X \subseteq A_\exists T^*_\exists X$.

From the above propositions on consistency we see: External admissibility ⇒ Internal admissibility ⇒ Basic Dungian admissibility, and Trust Closed admissibility ⇒ Trust Extended admissibility. Also, trivially, $\textbf{admissible}_X\, X \equiv \textbf{admissible}_D\, T^*_\exists X$. Moreover any Trust Extended Admissible set extends to a Trust Closed Admissible set.

PROPOSITION 15 *If X is a Trust Extended admissible set then it is a subset of some Y which is a Trust Closed admissible set, i.e. :*

$$\textbf{admissible}_X\, X \Rightarrow \exists\, Y \supseteq X.\, \textbf{admissible}_T\, Y.$$

5.10. Boolean Networks for Bipolar Argumentation

The boolean network could be obtained directly by compiling the semantics for the argumentation system. However, the presence of the transitive closure, T^*, in conditions presents an issue in that it requires considering connections between arguments at unbounded separation. Fortunately some simple transformation on the conditions avoids this problem.

PROPOSITION 16 *Trust Closed admissible is equivalent to $T_\exists X \subseteq X$ and* $\textbf{admissible}_D\, X$. *This follows immediately from $T_\exists X \subseteq X \Rightarrow T^*_\exists X \subseteq X$.*

PROPOSITION 17 *X is Externally Consistent is equivalent to there exists some Y subset of X such that Y is T-closed and conflict free.*

PROPOSITION 18 *X is Internally Consistent is equivalent to there exists some Y subset of X such that Y is T-closed and Y is not in conflict with X.*

So we may rewrite the axioms for the propositional system into a set of axioms involving the free variable Y, find models for the new system and extract the X component of the solution sets as positions of the bipolar systems.

5.11. The Other Side of Arguments

So far, setting aside the use of trust terminology, the development of argumentation theory in terms of Galois Connections has been in terms that should be recognisable to most argumentation theorists. But there is another way of interpreting these systems. We may consider the complement of the distrust relation and call the complement *implicit trust*. That is, if I do not *distrust* you, then I *implicitly trust* you. Such a change of perspective shifts the emphasis of the social relationships from an active ("I distrust") to the passive ("I implicitly trust"). The Bipolar systems are now systems with two positive relations, *implicit trust* and *explicit trust*. This change in perspective provides a useful tool in discussing social interaction such as (logical) reputation and privacy, and is accompanied by a matching change in the expression of the formal model. The change to the formal model is to translate the conditions on argument systems from the axiality $(R_{\exists}, R^{\forall})$, which express conditions in terms of a relation R, to conditions using the polarity (R_+, R^+)[3.90] which are expressed in terms of $\overline{R}$ using the identities between polarities and axialities[3.91, 3.92].

The Dungian conditions are readily re-expressed in terms of the complement of the attack relation and the new connection:

conflict free	$S \subseteq \overline{R}_+ S$
acceptable	$S \subseteq \overline{R}_+ \overline{R}_+ S$
complete	$\overline{R}_+ \overline{R}_+ S \subseteq S$
stable	$\overline{R}_+ S \subseteq S$

We may read the expressions $\overline{R}_+ X$ and $\overline{R}^+ Y$ as, "the set of things implicitly trusted by everything in X" and "the set of things that implicitly trust everything in Y", respectively. So the expression $Z \subseteq \overline{R}_+ X$ corresponds to "(everything in) Z is implicitly trusted by (everything in) X" and $Z \subseteq \overline{R}^+ Y$ corresponds to " (everything in) Z implicitly trusts (everything in) Y".

The internal and external consistency conditions for the standard bipolar systems may be similarly re-expressed using the complement of the distrust relation, implicit trust:

internally consistent	$S \subseteq \overline{A}_+ T^*_{\exists} S$
externally consistent	$T^*_{\exists} S \subseteq \overline{A}_+ T^*_{\exists} S$

That is, a set, S, is internally consistent if S is implicitly trusted by the trust closure of S and a set, S, is externally consistent if the trust closure of S is implicitly trusted by the trust closure of S (i.e. the trust closure is conflict free).

5.12. Trust Systems

Trust systems are a tool for formalising social trust, that is, formalising the acquisition of initial trust through referrals. A trust system can be thought of as defining a community of individuals that express a consistent view about judgements of trust and distrust. An individual has grounds to initially trust another if both the individual and the potential trustee both belong to such a community and there is a path of explicit referrals connecting the individual to the trustee. When these conditions arise between two individuals we say that they are *socially connected* by the community. Bipolar argumentation systems are used to model the notion of consistency of view with respect to trust and distrust judgements. Bipolar systems vary in the consistency conditions used. The consistency conditions of the standard system are "long range" because they use the transitive closure of the trust relation. While this is appropriate for pure argumentation it has untoward consequences when we consider trust and distrust as representing states of knowledge or belief of individuals.

Before continuing further let us introduce some terminology. Previously we called sets of arguments that satisfied the semantics *positions*. Now that we are switching to considering sets of individuals, let us call sets of individuals that satisfy the semantics *clubs* i.e. a club is a community that satisfies some contextually defined set of conditions.

Given a club, C, social connectedness via C can be formalised straightforwardly. First define T_C (the trust relation T relativised to the club C):

$$T_C \mathrel{\widehat{=}} \{(x, y) \in T \mid x \in C \wedge y \in C\}. \tag{5.37}$$

Social connectivity via C is now the relation T_C^*.

Different consistency conditions define different ideas of coherence of referrals made by a club. If we consider the standard bipolar system the consistency conditions express the following constraints on clubs:

- Simple consistency requires that clubs do not contain members that distrust one another.
- Internal consistency requires that clubs do not contain members that transitively trust individuals outside of the club that distrust some member of the club.
- External consistency requires that the club does not contain members that trust individuals outside the club that distrust one another.

One might say that these latter conditions are conditions on the reliability of judgements made by members of the club. Internal consistency requires that the judgements made by members about individuals outside the club align with one another. External consistency requires that transitive trust in individuals outside the club leads to a coherent policy of trusting beyond the club and that the club should be limited to a group that can make such coherent judgements.

However, these conditions are very strong in that they involve the consequences of following transitive relations (referrals). That is, an individual must be able to foresee conflicts arising arbitrarily far down the trust relation. From the point of view of individuals trusting, and distrusting, one another, this is too strong. So, in keeping with the idea that individuals have limited foresight, we consider an alternative notion of consistency which only requires avoiding immediate conflict conditions i.e. those conflicts arising after one step. Let us call the set $T_\exists X$ the (immediate) trust extension of X. Then we will say that a set X is *opinion consistent* if:

1. X is conflict free (i.e. $A_\exists X \subseteq \overline{X}$).
2. X does not attack its own trust extension (i.e. $A_\exists X \subseteq \overline{T_\exists X}$).
3. Nothing in the trust extension of X attacks X (i.e. $\breve{A}_\exists X \subseteq \overline{T_\exists X}$).
4. The trust extension of X is conflict free (i.e. $A_\exists T_\exists X \subseteq \overline{T_\exists X}$).

Or, adopting the phrasing of implicit trust:

1. The club implicitly trusts itself (i.e. $X \subseteq \overline{A}_+ X$).
2. The club implicitly trusts its own trust extension (i.e. $T_\exists X \subseteq \overline{A}_+ X$).
3. The club is implicitly trusted by its own trust extension (i.e. $T_\exists X \subseteq \overline{A}^+ X$).
4. The trust extension of the club implicitly trusts itself (i.e. $T_\exists X \subseteq \overline{A}_+ T_\exists X$).

If a club is opinion consistent then its members have a coherent view of how members and non-members are trusted: its members implicitly trust one another; if any member of the club explicitly trusts an individual outside the club then the rest of the club, at least implicitly, trusts the individual; any individual outside the club that is trusted by a member of the club implicitly trusts the whole club; if individuals outside the club are trusted by some member of the club they do not attack other individuals trusted by the club. In particular if the club does not have such coherency in its view of non-members one might reasonably call into question its view of members. For example, club members x and y both explicitly trust club member z but disagree over non-member w. In which case can we really be sure of their assessment of z?

Opinion consistency gives us a more bounded view of social connectedness. We will use the term society to mean simply a collection of interacting individuals. Given that we exist in a society, the question arises as to which opinion consistent clubs we belong to in that society and hence what social connections exist between ourselves and others.

Any opinion consistent club can be extended to one or more maximal opinion consistent clubs. Following the argumentation terminology we will refer to these as preferred clubs. Given that a pair of individuals may be socially connected in some such maximal clubs and not in others, the question arises as to how and when to form initial trust. If some club is preferred over others, e.g. members of my family, the friends I socialise with in the village pub, or the people I work with, then the process of determining the existence, or otherwise, of the social connection is straightforward. If, however, we have no preferred club in mind, then we have to decide how on aggregate to determine whether or not social connectivity exists.

Following the terminology introduced in the discussion of coherence in Chapter 4, we have the option of taking a sceptical stance, in which initial trust exists if all maximal clubs provide the required social connection; or a credulous stance, in which only some maximal clubs provide the required connection. That is, if we are very conservative, i.e. sceptical, in our trust stance, we will only extend initial trust to an individual who would be regarded as trustworthy in all the clubs we can construct from our society. If we are less conservative, i.e. credulous, we will extend trust to an individual with whom we have a social connection via at least one of the clubs we can construct from our society. We might feel that the credulous position is untenable as a basis for initial trust since there are both reasons to trust (social connections) and reasons not to trust (lack of social connections). However each club of which we are member is composed of individuals that we have no reason not to trust (by definition). In particular, we have no reason to distrust any of the individuals involved in the referral path connecting us to the individual with whom we are seeking to establish initial trust.

In practice we tend to use specific clubs for different purposes. I use the club composed of my peer group within the firm for which I work to resolve initial trust issues within the firm. I use the club of academic referral to resolve trust issues around published articles within my discipline. I use a club of friends whose opinion I respect for many day-to-day judgements for which I have no particular expertise.

Specific clubs may be formed for specific reasons. They may be bounded by social or contractual conditions that impose limits on the notion of trust beyond the club. For example, employees of a firm are expected not to discuss future business plans with individuals outside the firm[5]. In this context individuals outside the firm are regarded as uniformly distrusted by people inside the firm.

If there are no readily identifiable clubs, I look at the clubs constructible from a suitable slice of my society e.g. individuals whose opinions I have relied upon in matters with some relation to the issue at hand; or close friends who I believe would be concerned for my interests. Depending on the issue at hand I may seek simply to increase my confidence in the face of uncertainties, which may be improved by the presence of some social connection, or I may seek a high degree of assurance which can arise from all such clubs supporting a social connection.

[5] This in part applies because of insider trading legislation, as well as the general need for not leaking advanced information to the competition.

5.13. Certificate Authorities Again

In the previous chapter we considered the example of certificate authorities. We return to the example again to examine it from the point of view of social trust. Consider the Trust system in figure 5.10.

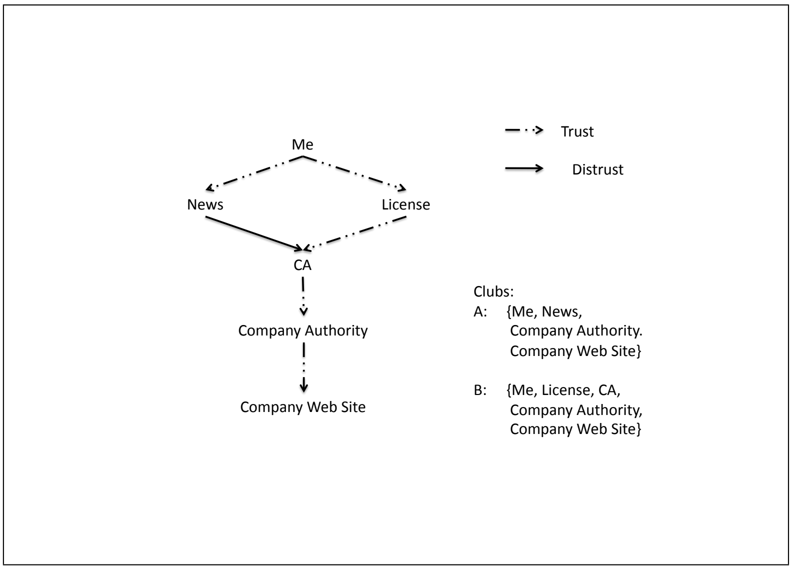
Me
News
License
CA
Company Authority
Company Web Site
Trust
Distrust
Clubs:
A: {Me, News, Company Authority. Company Web Site}
B: {Me, License, CA, Company Authority, Company Web Site}

Figure 5.10.: Single Certificate Authority

The individual “Me” trusts both “License”, the licensing body, and “News”, the source of news reports about the certificate authority. Computing opinion consistent maximal clubs[6] we see that, although we may implicitly trust the ‘Company Website”in both cases, that, if the “CA” is not a member of the club then there is no social connection between “Me” and the “Company Website”. Whether “CA” is a member of the club depends on whether we prefer “News” over “License” or vice versa.

Figure 5.11 illustrates the second case of the example with two Public Certificate Authorities and cross-signing.

Again, examining the maximal clubs, we see that in both of the maximal sets that both company websites are implicitly trusted. However, if one believes “News”, and so adopts the position set out by maximal set *B*, then there is no path between “Me” and either of the company websites in that set. So, in this context, no social connection exists and therefore there is insufficient grounds for social trust.

[6] We are interested in the existence of any opinion consistent club in which“Me” is socially connected to the “Company Website”. If they are socially connected it will be connected in a maximally consistent set.

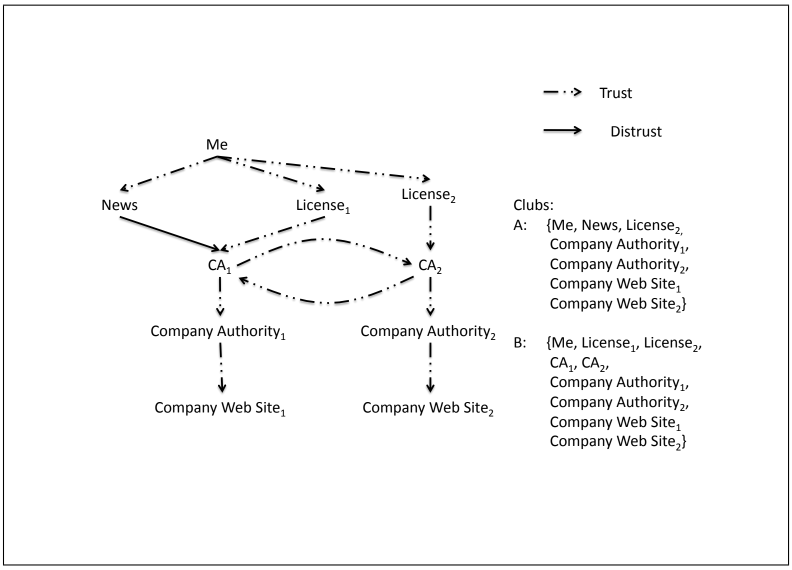
Trust
Distrust
Me
News
License1
License2
CA1
CA2
Company Authority1
Company Authority2
Company Web Site1
Company Web Site2
Clubs:
A: {Me, News, License2, Company Authority1, Company Authority2, Company Web Site1 Company Web Site2}
B: {Me, License1, License2, CA1, CA2, Company Authority1, Company Authority2, Company Web Site1 Company Web Site2}

Figure 5.11.: Cross-signing

5.14. Privacy: An Application of Argumentation and Trust

Privacy is important to us. It is a shield that we stand behind. It protects us from exploitation and manipulation by those who do not have our interests at heart.

What makes information *private* information is that it is information about *us* that we feel we have the right to control, or it is about *us* and *others* and we jointly feel we have the right to control; or, to put it rather more bluntly, private information is information which we feel it is *nobody else's business* to know. What legitimises this notion of privacy is the general consensus of society that such-and-such information is, indeed, *nobody else's business*, and that the acquisition of, or general publishing of, such information by another party would, in one way or another, diminish the lives of those concerned. There is a moral dimension to the judgement by society that information is private. Privacy cannot legitimately be invoked to hide moral turpitude or hide a crime. Since we may clearly attempt to place information into this category of privacy that does not belong there, society concedes that there is a notion of *public interest* that overrides considerations of privacy.

Keeping information private reduces the scope for intrusion and interference in our lives. Privacy of information permits us to lead a life of relative freedom. And yet we must balance privacy against necessary disclosure which permits others to act on our behalf and in our interest.

When we disclose information that we regard as private, it is to enable another individual to act as *our* agent, making decisions in *our* stead and in *our* interest. Our trust is that the agent will not use this information to his advantage and our disadvantage.

Privacy is contextual and governed by a contextually dependent collection of *social norms* that define how particular kinds of information *should, and should not,* be handled in each social context. Generally we disclose information to people provided that we believe that they will follow the appropriate social norms for the context and kind of information.

The undertaking to maintain privacy is a promise, implicit or explicit, not to disclose information without need (the promise of *need*); not to disclose information to those who will not maintain the social norms governing the privacy of the particular kind of information disclosed (the promise of *safety*) and the promise of only disclosing the information for the purpose for which it was originally provided (the promise of *purpose*).

Rarely is disclosing private information a matter of sharing a secret with another individual. Often the information is not secret in that no particular measures have been taken to protect it, and it is often discoverable by other means. Also it is rare that the information is to be shared with an individual. Usually disclosure means that some group of individuals, or some organisation, will have access to the information for some specified purpose. Let us call such a group or organisation a *community*. Communities come equipped with *practices*, *procedures* or *processes* that cause information to be transferred between individuals. We will call these *community processes*, or simply *processes* for short. Processes represent the known modes of information transfer within the community. For example, an insurance firm may have a procedure for handling a claim that requires the claim to be passed between individuals in certain roles, e.g. a Loss Assessor and a Payment Authority. But we should be aware that there are also be other processes, including accidental transfer e.g. a document left on a shared printer or an overheard office conversation. We may think of these as official community processes and unofficial community processes. In discussing privacy we must be aware of these unofficial community processes lest we deceive ourselves into a misplaced feeling of safety.

When disclosing information to a community in the first instance we often disclose information to an individual (as opposed to in a broadcast to the whole community). In such cases part of the trust that is placed in the initial recipient of the information is that, even within the community, the information will not be shared without need. Indeed, even within the community, we reasonably expect the three promises of *need*, *safety* and *purpose* to be met as information is passed around the community. And one reason we may decide not to disclose information is that, although we believe the primary recipient will follow the appropriate norms, we do not share the primary recipient's beliefs about various third parties to whom the primary recipient may further disclose the information. We may think of this as assessing sources of potential leakage.

This leads us to two ways of assessing the trustworthiness of a community in handling private information. We may assess this trustworthiness by the referral process we discussed above. But we may also, if we have access to the appropriate information, assess how information might flow within the community. We expect information to be passed between individuals who are socially connected by explicit trust relationships. If we have a map of the referral process within the community then we may predict which individuals in the community may end up in possession of the information and determine whether or not we feel that this is acceptable. More precisely we require that we at least implicitly trust every possible recipient of our private information. One might consider these two approaches as “black box” analysis of trustworthiness and “white box” analysis of trustworthiness in handling private information.

The white box condition is simply formalised as, if x discloses to y and y belongs to club C then:

$$(T_C^*)_\exists\{y\} \subseteq \overline{A}_+\{x\} \tag{5.38}$$

Analogous to the problem of deciding to whom we should disclose information is the problem of who should be granted access to information by some access control system. Here policy fulfils the role of social norms and the leakage condition corresponds to the notion that even if an individual has, as an individual, appropriate access privileges, they may still not be granted access if there are potential leakage routes from that individual to other, less privileged individuals. For example, consider an individual who shares an office with some less privileged individual. If information is disclosed to the privileged individual it is not unreasonable to suspect that some information may leak to the less privileged individual. In judging whether or not to grant access to the information we may consider consulting both a clearance database and an office accommodation database to decide whether or not to grant access (or to determine what form of access to grant e.g. access granted in a secure reading room).

To set up the analogy formally we introduce a prototype individual at each clearance level, say c_l for each level l and an implicit trust relation C between c_l and all individuals cleared at level l (or, if we wish, l or higher). We will treat the organisation as an opinion consistent club and people who share offices will be regarded as having explicit trust in one another (there may be other explicit trust relationships). The explicit trust relations will be called E. To disclose information at level l to an individual i we require both i to be cleared to level l, and everyone explicitly trusted by i in the organisation is cleared to level l (or above), i.e. $(E^*)_\exists\{i\} \subseteq C_+\{c_l\}$. In the terminology introduced earlier we have an official community process of disclosing and an unofficial community process of (potential) information transfer by office sharing. We may contrast this discussion of privacy with the discussion of confidentiality[7] in computer system security. The difference is one of perspective. With confidentiality we tend to concern ourselves with the *technical structure* of the non-disclosure/non-leakage condition. With privacy, at least as conceived here, we are concerned with the social processes around disclosures.

5.15. Conclusion

This chapter has created an approach to reasoning about Social Trust based on adapting the techniques of argumentation theory. The essence of the adaptation has been to switch from considering how abstract arguments attack and support one another to considering how individuals in a community attack and support one another in the matter of trust. The resulting theory has then been applied to the

[7] See appendix C for a discussion information security using Galois Connections

analysis of trust placed in certificates authorities within Public Key Infrastructures and in the role of trust in privacy. The underlying reasoning technique in this and the previous chapter is propositional model based reasoning. The next chapter explores the connection between the techniques used in these two chapters.

6. Coherence and Argumentation Theory

In which is discussed:

- *How Preferential Coherence is connected to Argumentation Theory.*
- *The semantics of Coalition argumentation systems.*

6.1. Introduction

The principal difference between the techniques of Preferential Coherence and argumentation theory is that preferential coherence deals with the case of conflict between two sets of sources, whereas argumentation theory only deals with sources being in pairwise conflict. In other respects the systems are remarkably similar and, within the limits circumscribed by the 'set' versus 'pairwise' interaction, the systems can simulate one another. This should not be particularly surprising given that both are extensions of reasoning by classical consistency checking.

In this chapter we examine some of the connections between Preferential Coherence and argumentation theory. We consider both how maximally consistent sets of labelled propositions are connected to argumentation theory, and how maximally preferred sets of propositions are connected to argumentation theory. The examination of maximally consistent sets leads to the consideration of an extended class of Dungian argumentation systems in which the attack relation is between sets of arguments rather than individual arguments.

6.2. Argumentation with a Symmetric Attack Relation

First let us consider a Dungian argumentation system $(\mathbf{Args}, R)$ with R symmetric, i.e. $R = \breve{R}$. Let S be a conflict free set. Immediately we see S is acceptable since $\breve{R}_\exists S = R_\exists S$. Moreover, this also means S is complete. If S is maximally consistent then S is preferred and since R is symmetric and S maximal, everything not in S is attacked by S and so S is stable (i.e. $S \cup R_\exists S = \mathbf{Args}$).

Examples of such argumentation systems can be constructed using the labelled propositional system introduced in Chapter 4. Let *sources* be a set of labels and *propositions* be a set of propositional letters in one-to-one correspondence with the sources. If **A** is a source name, let A be the corresponding propositional letter. Let **Args** be a set of observations (a set of labelled propositions), containing:

- For each source **A**, the assertion **A** : A;
- For each source **A**, and each source **B**, optionally **A** : $\neg B$;

- and no other assertions.

A Dungian argumentation system is constructed by taking the sources as a set of arguments, **Args**, and the (symmetric) attack relation, R, as **A** attacks **B** if, and only if, $\mathbf{incons_{Args}}\{\mathbf{A}, \mathbf{B}\}$. Trivially, the maximally consistent theories for this construction are the preferentially coherent theories obtained with the empty order relation and coincide with stable positions in the argumentation theory semantics.

6.3. Breaking Symmetry

Continuing with the example of propositional systems, we next consider what happens when we break the symmetry of the attack relation by introducing an ordering between sources. Attacks will be oriented from the most preferred to the least preferred source. If no preference is expressed between **A** and **B**, then they will be regarded as mutually attacking. That is **A** attacks **B** if, and only if, the sources are inconsistent with one another and **A** is preferred to **B** or there is no expressed preference between **A** and **B**.

If we examine an arbitrary maximally consistent set, say S, under this additional condition, we may note that an item may be in $\overline{S}$ either because S attacks it or it attacks S. In the argumentation theory semantics, the acceptability condition guarantees that a conflict free set counter attacks all its attackers but leaves open the possibility of self inconsistent elements being present (these cannot be in S but do not necessarily attack S). In the propositional semantics, such a self inconsistent element, e.g. $\{\mathbf{A} : A, \mathbf{A}, \neg\, A\}$, is inconsistent with every other element and so either attacks, or is attacked by, every other element (or both), depending on the preference order between elements. So, if acceptability holds for S (that is $\breve{R}_{\exists}S \subseteq R_{\exists}S$), then S is stable.

Continuing the example, if we now require further that: $\mathbf{A} \succ \mathbf{B}$ if, and only if, **A** asserts $\neg\, B$, then there is a correspondence between the preferential coherence semantics and the Dungian semantics. A solution set S under the preferential coherence semantics is also a position under the stable semantics for the corresponding argumentation system.

By definition, a preferentially coherent set is maximally consistent and anomaly free. A maximal consistent set under the propositional semantics is clearly a maximal conflict free set in the argumentation semantics. We now show that anomaly freeness is sufficient to guarantee stability of the corresponding argumentation system (and by proposition 5 is therefore acceptable and complete).

PROPOSITION 19 *If the set S is anomaly free then each point in $\overline{S}$ is below some point in S (i.e. $\overline{S} \subseteq (\succ)_{\exists}S$).*

def $S \sqsupseteq \overline{S}$

$\equiv\ \forall x \in S.\, \exists y \in \overline{S}. y \succeq x$

$\equiv\ \overline{S} \subseteq (\succeq)_{\exists}S$

$\equiv\ \overline{S} \subseteq (\succ)_{\exists}S \cup (=)_{\exists}S$

$\equiv\ \overline{S} \subseteq (\succ)_{\exists}S \cup S$

hint $\overline{S} \cap S = \emptyset$

$\equiv\ \overline{S} \subseteq (\succ)_{\exists}S$

PROPOSITION 20 *S is stable.*
By construction the attack relation is $\succ$ and by proposition 19: $\overline{S} \subseteq (\succ)_{\exists}S$.

We can also simulate preferential and stable semantics for argumentation systems in the coherence semantics. Indeed we have already done most of the work in giving the Boolean Network translation of argumentation systems by:

- taking the set of sources to be the set of arguments, **Args**, of the argumentation system;
- taking the set of propositional letters to be in one-to-one correspondence with the set of labels, so that if **A** is a source then A is the corresponding propositional letter;
- for each source **A** making the assertion **A** : A;
- for each source **A** making the assertion **A** : θ, where θ is the whole of the Boolean Network translation for the argumentation system;
- taking the universal relation[1] as the preference relation.

If θ is the translation for admissible or complete semantics then the solution sets of the coherence system correspond to preferred positions, whereas if the translation is for stable semantics then the solutions sets correspond to stable positions. This is because preferential coherence finds the maximal consistent sets and then filters these by the conditions on the ordering relation. By taking the ordering relation as the universal relation this filtering is trivialised leaving the maximally consistent sets.

6.4. Discussion

The above correspondence shows that when we limit consideration to labelled propositional systems in which all conflicts arise directly between pairs of sources then the preferential coherence semantics can be mimicked by an argumentation system, and any argumentation system with preferential or stable semantics can be mimicked by a preferential coherence system.

This begs the question of whether there is an argumentation system with a more general correspondence to preferential coherence? The essential problem with argumentation systems is that conflict is between pairs of arguments, hence situations such as a three way mutual incompatibility cannot be represented. The simplest solution to this seems to be to abstract directly over the labelled

[1] A.K.A. Chaos.

propositional systems to create a new form of argumentation system that deals with the situation directly.

In labelled propositional systems sets of labels are connected by a symmetric inconsistency relation. A set of labels is consistent if none of its subsets are inconsistent with any other (i.e. if its subsets are only inconsistent with sets that are not its subsets). A set of labels is maximally consistent if it is consistent and it is inconsistent with every other set (i.e. its subsets are inconsistent with every set that is not a subset of itself).

We define a symmetric coalition system as a pair $(\mathbf{Args}, \not\leftrightarrow)$ where **Args** is a set of arguments and $\not\leftrightarrow$ is a symmetric relation over sets of **Args**, $\mathbb{P}(\mathbf{Args})$.

- *Consistent*: A subset S of **Args** is consistent if, and only if,

$$(\not\leftrightarrow)_{\exists}\, \mathbb{P}\, S \subseteq \overline{\mathbb{P}\, S};$$

- *Complete*: A subset S of **Args** is complete if, and only if,

$$\overline{\mathbb{P}\, S} \subseteq (\not\leftrightarrow)_{\exists}\, \mathbb{P}\, S;$$

- *Maximally Consistent*: A subset S of **Args** is maximally consistent if, and only if, it is consistent and complete.[2]

Next we extend symmetric coalition systems to preferential coalition systems by adding an order relation, $\succcurlyeq$, over subsets of **Args**. A preferential coalition system is a triple $(\mathbf{Args}, \not\leftrightarrow, \succcurlyeq)$ such that $(\mathbf{Args}, \not\leftrightarrow)$ is a symmetric coalition system and $\succcurlyeq$ is a binary relation over subsets of **Args**:

- *preferentially acceptable*: A subset S of **Args** is preferentially acceptable if, and only if, $(\breve{\succcurlyeq})_{\exists}\{S\} \subseteq (\succcurlyeq)_{\exists}\{S\}$;
- *preferentially stable*: A subset S of **Args** is preferentially stable if, and only if, $\{\overline{S}\} \subseteq (\succcurlyeq)_{\exists}\{S\}$.

A preferential coalition system has preferential coherence semantics when its solution sets are required to be maximally consistent, preferentially acceptable and preferentially stable. It is not difficult to see that if we take a preferential coherence system with set of observations *Obs* and translate it into a preferential coalition system by taking the consistency relation between sets of labels to be

[2] These conditions can equally be phrased in terms of the consistency relation between sets $\leftrightarrow \mathrel{\hat{=}} \overline{\not\leftrightarrow}$ and the $\_^{+}$ Galois operator. They become:

- *Consistent*: A subset S of **Args** is consistent if, and only if, $\mathbb{P}\, S \subseteq (\leftrightarrow)^{+}\, \mathbb{P}\, S$, i.e. the set consistent with every member of $\mathbb{P}\, S$ includes the members of $\mathbb{P}\, S$;
- *Complete*: A subset S of **Args** is complete if, and only if, $(\leftrightarrow)^{+}\, \mathbb{P}\, S \subseteq \mathbb{P}\, S$, i.e. $\mathbb{P}\, S$ contains all the sets consistent with every member of $\mathbb{P}\, S$;
- *Maximally Consistent*: A subset S of **Args** is maximally consistent if, and only if, it is consistent and complete, i.e. the sets that are consistent with every member of $\mathbb{P}\, S$ are exactly the sets in $\mathbb{P}\, S$.

$A \not\rightarrowtail B \mathrel{\widehat{=}} \mathbf{incons}_{Obs}(A \cup B)$ and order relation $\sqsupseteq$ then the two systems have the same solution sets.

Although this gives an operator semantics, unfortunately, the presence of the unit set operator, $\{\_\}$, and power set operator, $\mathbb{P}\,\_$, mean that we cannot evaluate the system using the Boolean Network translation.

If we consider symmetric coalition systems and consider going in the other direction i.e. from a symmetric coalition system to a labelled propositional system, then it is clear that not all systems will correspond to labelled propositional systems because the inconsistency relation of the labelled propositional system is both left and right monotonic i.e. where for a binary relation over sets R we define:

- *left monotonic*: if whenever xRy and $x \subseteq x'$ then $x'Ry$;
- *right monotonic*: if whenever xRy and $y \subseteq y'$ then xRy';

which we might equivalently express as:

- *left monotonic*: $R_{\exists} \circ (\supseteq)_{\exists} = R_{\exists}$;
- *right monotonic*: $(\subseteq)_{\exists} \circ R_{\exists} = R_{\exists}$.

If we require the symmetric coalition systems to be both left and right monotonic then we can interpret the system by a labelled propositional system by a coding similar to the above Boolean Network coding i.e.

- taking the set of sources to be the set of arguments, **Args**, of the argumentation system;
- taking the set of propositional letters to be in one-to-one correspondence with the set of labels, so that if **A** is a source then A is the corresponding propositional letter;
- for each source **A** making the assertion **A** : A;
- for each source **A** making the assertion **A** : θ, where θ is the attack relation given as $A_1 \wedge \ldots \wedge A_n \equiv \neg\ (B_1 \wedge \ldots \wedge B_m)$ for each attack $\{A_1, \ldots, A_n\} \not\rightarrowtail \{B_1, \ldots, B_m\}$;
- taking the universal relation as the preference relation.

For this reason we call the left and right monotonic symmetric coalition systems, the symmetric conjunctive coalition systems.

6.5. Another Approach to Argumentation

It would be clearly useful to have a form of argumentation theory that could address amalgamation of information. Here we briefly consider a possible candidate for such an approach to argumentation.

Taking a purely formal approach to the problem we extend Dungian semantics to deal with sets of arguments i.e. the attack relation is over sets of arguments rather than individual arguments. We then require the conditions:

- collections of arguments can collectively attack a collection of arguments
- a solution set, X, is a set of arguments
  - such that no subset of X attacks a subset of X
  - such that any set of arguments, Z, which attacks a subset of X is itself attacked by some subset of X

If the attack relation is $R : \mathbb{P}\,\mathbf{Args} \leftrightarrow \mathbb{P}\,\mathbf{Args}$ then we are seeking solutions X such that:

$$R_\exists Y \subseteq \overline{Y} \tag{6.1}$$

$$\breve{R}_\exists Y \subseteq R_\exists Y \tag{6.2}$$

and such that

$$X = \bigcup Y. \tag{6.3}$$

It is easily seen that if the first condition is to guarantee that no subset of X is attacked by a subset of X then Y must be $\mathbb{P}\,X$. Similarly, if the second condition is to guarantee that X defends itself then again Y must be $\mathbb{P}\,X$. That is we seek solutions X such that:

$$R_\exists(\mathbb{P}\,X) \subseteq \overline{(\mathbb{P}\,X)} \tag{6.4}$$

$$\breve{R}_\exists(\mathbb{P}\,X) \subseteq R_\exists(\mathbb{P}\,X) \tag{6.5}$$

We might call this the *free* coalition system in that no constraints hold between sets of arguments. In particular there are no monotonicity conditions, meaning, for example, that $\{a\}$ may attack $\{b\}$, but $\{a, c\}$ may not attack $\{b\}$, or more concretely, "if it is hot, I will not go running", and "if it is raining, I will not go running" does not necessarily entail, "if it is raining and it is hot, I will not go running"[3].

However, if we decide that there are additional internal consistencies that should hold between sets of arguments then these are representable as constraints between propositional variables representing coalitions. The extreme case is to require the argumentation to be monotonic i.e. that the attack relation is closed under super sets on the left and the right. That is, let $R^{\#} \mathrel{\widehat{=}} \supseteq \fatsemi R \fatsemi \subseteq$, then R is left and right monotonic if, and only if, $R = R^{\#}$, i.e. if $\{a\}$ attacks $\{b\}$ then $\{a, c\}$ attacks $\{b\}$, and also, $\{a\}$ attacks $\{b, c\}$. In between these free semantics and the monotonic semantics there are many other possibilities. For example, we may take the attack relation as left monotonic but not right monotonic, which might be interpreted as additional information may rebut an attack. Similarly we may consider an attack relation which is right monotonic but not left monotonic, or

[3] This example is due to Professor Tim Norman, who claims not to run when it is hot, or when it is raining, but has gone running in Hawaii when it was both hot and raining.

adopt other, more ad hoc, relations between antecedent sets and conclusion sets of the attacks relation.

Although we do not explore this topic further here, the non-monotonic argumentation system offers interesting possibilities when we consider trust and distrust relations. For example, if we think of attacks in terms of a distrust relation, then "Jill distrusts John" and "Jill distrusts Jeremy" does not imply "Jill distrusts John and Jeremy" (for example, Jill's distrust may arise from concerns over competency in a particular area of expertise but John and Jeremy's competencies may be complementary, so together they are trusted). This requires at least right non-monotonicity.

To evaluate a free system we can use similar techniques to those used in Chapter 5 to evaluate such a system. Clearly, if sets of arguments are considered as fully independent of one another then we may represent these sets by propositional variables and follow much the same scheme as before. That is, we may take a set of $2^{|\mathbf{Args}|}$ propositional variables to represent all subsets of arguments (aka Coalitions) and define:

$$\bigwedge_{x \in \text{Coalitions}} \Big(\bigvee_{(y,x) \in R} \mathcal{X}_y \Rightarrow \neg\, \mathcal{X}_x \Big) \tag{6.6}$$

$$\bigwedge_{x \in \text{Coalitions}} \Big(\mathcal{X}_x \Rightarrow \bigwedge_{(y,x) \in R} \bigvee_{(z,y) \in R} \mathcal{X}_z \Big) \tag{6.7}$$

i.e. we just have a version of our earlier semantics using a bigger set of propositional variables.

Additional constraints, such as monotonicity, expand the attack relation. In the case of monotonicity the expansion significantly increases the number of clauses in the boolean expression used for model finding. However many of these additional clauses, at least in this specific case, are redundant, which raises the possibility of seeking minimal codings, which, however, must be left as a topic for future research.

6.6. Conclusion

This chapter has explored the connection between the techniques of propositional model based reasoning used in Chapters 4 and 5, i.e. the techniques for reasoning about Knowledge on Trust and Social Trust. As a result of the analysis, an approach to argumentation has been proposed that generalises Dungian argumentation systems from attack relations between arguments, to attack relations between sets of arguments. The resulting *free* system is non-monotonic; however, the system permits the addition of further constraints on attacks between sets of arguments to recover monotonicity if required. Finding solutions by propositional model finding is still possible in such a system but the possibilities for compact coding of monotonicity constraints as appropriate axioms, or axiom schemes, still needs to be investigated. That said, the generalised Dungian system suggests itself as a framework within which the reasoning techniques for Knowledge on Trust and

Social Trust might be unified in future work. This subject is further discussed in the next chapter (see section 7.3).

The next chapter examines a number of topics around Knowledge on Trust and Social Trust and indicates possible future directions for their development.

7. Discussion

In which is discussed:

- *The barriers to Application.*
- *Deduction Systems for Generalised Argumentation.*
- *The Risk of Trust.*
- *Making Promises Formal.*
- *Trust and Balance.*
- *The Dynamics of Trust.*
- *Social Science and Signed Graphs.*

7.1. Introduction

In this chapter, before a final conclusion, we turn to examine the prospects for a few of the possible topics for further development. First we briefly examine some of the issues around deploying the concepts as a technology. Next we turn to examine one of the directions in which we might take the logic, that of proof systems for generalised argumentation systems. This is followed by a brief consideration of one of the connections between trust as a logical notion and the *risk* associated with the act of trust, which is followed by a discussion of work involved in formalising the logic of promises. We then briefly turn to the discussion of the notion of the dynamics of trust and an outline of a lately realised connection with mathematical sociology.

7.2. Prospects for Application

One area where we might expect to seek application for reasoning about trust is the Web, and, perhaps even more so, the nascent Semantic Web. Today, web security is based on preserving the integrity of trust decisions made by the user. Unfortunately, users are often ill equipped to make informed decisions, lacking both the technical expertise to truly understand what is being asked of them, and the specific, contextual, knowledge required to make an informed decision. Asking the user is, in practice, simply a way for a system provider to evade responsibility for subsequent ills a user may suffer. It is tempting to provide a better solution by providing some form of trust reasoning that can either make the decision on behalf of the user, or, at the very least, provide a user with options on how to proceed, backed by rational, and understandable, arguments. Desirable as this may be there are never the less real barriers to the adoption of such a strategy which must be considered.

There are essentially four problems: Coding, Complexity, Disclosure and Liability.

Coding: Logic applies to formal systems, that is, to systems where a sufficient level of coding has taken place to reduce the semantic problem to a syntactic one. Unfortunately, although there are vast resources of information on the Web, the majority of it is "uncoded" and requires significant effort to formalise both it and its context.

Complexity: Even when we deal with finite cases, logical methods usually do not scale well. For example, Thomas Schaefer showed that there are exactly six forms of propositional theory whose satisfaction problem is polynomial time decidable, all others being NP[1][100]. Clearly Preferential Coherence is NP and it is not difficult to see that, in general, a generalised argumentation system will be NP. Indeed one can show that even the basic Dungian system is NP complete by embedding the satisfaction problem for propositional logic[2].

These first two problems are quite general and apply to a wide range of possible uses of semantic processing on the web and are potential barriers to the development of the Semantic Web[3,4].

[1] Briefly, Schaefer defines a logical relation as a relation over $\{0,1\}^k$ for $k \geq 1$ and, given a set of logical relations $S = \{R_1, \ldots, R_n\}$, defines an S-formula to be a conjunction of clauses of the form $R(\eta_1, \ldots, \eta_k)$ where $\eta, \ldots, \eta_k$ are variables. The S-satisfiability problem is then defined as the problem of deciding whether a particular S-formula is satisfiable. The set of all satisfiable S-formulas is called SAT(S). Schaefer then shows the class SAT(S) to be polynomial time satisfiable if, and only if, the class S is such that:

1. Every relation in S is satisfied when all variables are 0;
2. Every relation in S is satisfied when all variables are 1;
3. Every relation in S is definable by a CNF formula in which each conjunct has at most one negated variable;
4. Every relation in S is definable by a CNF formula in which each conjunct has at most one un-negated variable;
5. Every relation in S is definable by a CNF formula having at most 2 literals in each conjunct;
6. Every relation in S is the set of solutions of a system of linear equations over the two element field $\{0,1\}$.

Otherwise the class is NP time satisfiable.

[2] Briefly define the system:

- The set of arguments is $a_1, \ldots, a_n, a'_1, \ldots, a'_n, c_1, \ldots, c_m$ and d.
- For each a_i, a_i attacks a'_i and a'_i attacks a_i.
- Given a satisfaction problem in conjunctive normal form in variables $a_1, \ldots, a_n$, define for each clause $C_j = \eta_1 \vee \ldots \vee \eta_k$ of the CNF, if $\eta_i = a_l$ then there is are mutual attacks betweena_i and c_j and if $\eta_i = a'_l$ then there is are mutual attacks between a'_l and c_j.
- For each c_j there are mutual attacks between c_j and d.

After simplification we can derive $d = \bigwedge_k \bigvee_i \eta_i^k$ where each η_i^k is either a_i^k or a'^k_i. The satisfaction problem is then equivalent to asking the question: "Is there an admissible set that contains d". That is, if we have an algorithm to find all admissible sets, then by inspecting these sets we have an algorithm to solve the satisfaction problem.

[3] Particularly when one reflects that some proposed Semantic Web tools include languages based on Description Logic which is almost identical to the language of generalised argumentation systems.

[4] These are not the only form of barrier to the development of the semantic web. Another is general

The coding problem can be addressed in one of two ways. We may either restrict applications to areas which require little coding, or we may hope for at least the partial success of the semantic Web project to provide suitably coded domains. Likewise the complexity constraint may be approached in one of two ways. We may either limit the size of problem that we address to suitably small domains (where small may be of the order of a hundred propositional variables with reasonable time limits and the right sat-solver), or we might adopt an approximation approach to sat-solving, in which we accept that solution sets have an error probability but at the gain of getting a polynomial time approximate solution.

Disclosure: Are people are willing to make the appropriate information available (and are they are willing to be associated with that information)? It is not hard to imagine circumstances in which making one's relative assessments of trustworthiness available for public scrutiny could lead to unpleasant consequences. It is probable that if we disclose such information it is likely to be on a limited basis (e.g. to some of those we trust), or done anonymously. If the latter then there is the problem of the honesty of the disclosure. The three circumstances in which disclosure might be reasonable are when the information is expected to be truly private, e.g. I build a trust model on my computer that is used by my browser to make decisions on my behalf; when the disclosing body is openly fulfilling self interest by disclosing information, e.g. Barbour's site, `www.barbour.com/counterfeit-education`, publishes a list of url's of companies that sell fake Barbour merchandise; and when the disclosing body is acting openly and in the public interest (but possibly as a commercial venture) e.g. publishing reliable url blacklists.

Liability: Where does the liability lay for bad decisions? This is particularly a problem where we deal with an algorithm and multiple sources of information. In practice, where there is no issue of fraud or other criminal intent, we might assume that a trust reasoner is a "reasonable effort" tool in the sense that it takes all reasonable precautions in reaching a conclusion from the information available, and that information suppliers take reasonable precautions to ensure the appropriateness of their information[5] (this is in explicit contrast with a licensing

economic considerations. Although the Semantic Web may offer significant benefits if developed, it may not offer the necessary incremental incentives to fuel its progressive development. A major part of the success of the Web was due to its openness. But an even greater part was due to the advantages that the Web offered businesses and individuals, which was primarily due to it being an effective publishing medium allowing it to become an information and catalogue shopping phenomenon. Individuals, businesses and non-profit organisations could immediately see how to take advantage of the new medium. There seems to be a lack of this second factor for the Semantic Web. The Semantic Web is not rapidly colonising our meme space as did the web. Industry is not rushing to put semantic coding into sites, nor are individuals finding they need to reach out to interact semantically with web agents. Though there are potential rewards to be gained from the Semantic Web, they appear to be rewards that cannot be reaped by progressive investment and evolution and therefore are not susceptible to the same kind of evolution as took place with the Web. A second point that may restrict its development is that, unlike the web, it has been conceived of as a totality rather than as a work in progress. The web grew from many independent efforts which both provided an enabling platform for the first http/html protocol exchanges and subsequently as the repertoire of protocols and languages were revised and expanded under a conjunction of independent experimentation and advantage seeking. These factors seem missing from the world of the Semantic Web.

[5] Here we merely say appropriate because some information may well be opinion rather than fact. This is appropriate provided it is portrayed as opinion rather than fact. Opinion masquerading

arrangement that accepts no liability for the conclusions of the reasoning). Phrasing this slightly differently, a user may expect the trust reasoner to draw reasonable conclusions from the data available and neither tool provider, nor the information providers have liability if this is the case. Without such a commitment the user has no reason to accept the conclusions of the tool nor have the providers reason to accept the risk of providing either information or tool. However, this itself needs to be underwritten by a social or legal framework which ensures that all parties meet their obligations.

If we can overcome each of these problems then one might build useful trust enhancing technology for the Web. An example of such a technology would be a trust service that provides trust advice about websites that uses a limited number of information sources. These sources may cross comment on the trustworthiness of one another and upon other web sites. Complexity is a major issue in such an application. One route out of this particular problem is for the user to nominate particular trust communities, i.e. clubs, that form an acceptable basis for social connection. The computational problem then reduces to (a) checking that the club conditions are maintained, and (b) checking there is a social connection within the club. Significant work is required to explore the feasibility of such an approach.

7.3. Prospects for Direct Proof Systems for Generalised Argumentation Systems

The logic of argumentation has been portrayed as a process of finding possible positions against a given semantics. However the Boolean Network translations can be viewed in a different light. Given a finite generalised argumentation system, $(\mathbf{Args}, R_1, \ldots, R_n)$, the Boolean Network translation into a corresponding propositional system (P, Ax) provides an axiomatisation of the argumentation system, provided that minimisation/maximisation is not used in the semantics, and may be used to prove that various properties follow from the system. This raises two interesting questions. Firstly can we produce some direct reasoning system for generalised argumentation systems when minimisation/maximisation is not used, and secondly, can we produce a direct reasoning system when it is used? Here we will briefly outline a system that addresses the first question and speculate on how we might address the second.

(There are, of course, a number of other alternative approaches one could consider. The two most obvious are either to adapt Brink's algebra of Boolean Modules[21][6], or to take an approach based on a (multi) modal logic view of

as fact is inappropriate.

[6] A Boolean Module combines a Boolean Algebra, which, in our application, we may think of as capturing the idea of sets of arguments, with a relation algebra, which we may think of as capturing the various attack and support relations. These algebras are joined to one another by a product operation that applies a 'relation' to a 'set' (i.e. an element of the relation algebra to an element of the set algebra). Brink defines a left Boolean Module by taking this product to be the Pierce Product (co-image) and the right Boolean Module by taking this product to be the image operator. Since the relational operators include *converse* these two operators are inter-definable.

A boolean algebra is always isomorphic to a field of sets and so boolean algebras can be said to precisely capture the notion of intersection, union and complementation over a field of sets.

argumentation systems in which each relation corresponds to a new modality, in essence generalising Grossi's work on Dungian semantics [44].)

The relational language for generalised argumentation systems can be viewed as expressing a particular fragment of first order logic in which unary predicates play the role of sets and binary predicates play the role of relations (after all, this is their interpretation in the model theory of first order logic). Solution sets are therefore modelled as unary predicates which, given a definition of the set of arguments and the relations of the argumentation system, make the semantic conditions true. This approach is similar to Ewa Orlowska's approach to relational systems based on Dual Tableaux [85]. We will briefly elaborate the idea assuming some familiarity with Genzen's two-sided sequent formalism (see, for example Goran Sundholm's, "Systems of Deduction" in Handbook of Philosophical Logic. Vol. I [111], or the translations of Gerhard Gentzen's original papers by Szabo in "The collected papers of Gerhard Gentzen" [112]).

Given a generalised argumentation system, $(\mathbf{Args}, R_1, \ldots, R_n)$, with semantic conditions, $\mathbf{Sem}(X_1, \ldots, X_m)$[7], axiomatised by the set of axioms Ax_{arg}, then $S_1, \ldots, S_m$, axiomatised by the axioms Ax_{Sol}, is a solution if, and only if, Ax_{args}, Ax_{Sol} $\vdash C^T$, where C^T is the first order translation of each condition C in $\mathbf{Sem}(S_1, \ldots, S_m)$. This translation is given in figure 7.1. Note that Ax_{args} contains the axiom $\forall x.x \in \mathbf{Args}$ to ensure that every term corresponds to an argument.

It is however generally desirable to reason directly in the language of generalised argumentation systems, or in first order logic augmented with this language, rather than detouring through a translation. Given the first order translation it is possible to create derived rules of inference for each operator in the language of generalised argumentation systems. This will be briefly illustrated by giving two sided Gentzen rules for the language[8]. In order to allow the use of relation symbols as abbreviations, and to reason about particular finite systems in the language, the language is extended with relational equality and finite explicit displays of sets, binary relation and constants, so that, for example, $\{a_1, \ldots, a_n\}$ is an explicit set of n constants and $\{a_1 \mapsto b_1, \ldots, a_n \ \mapsto b_n\}$ is an explicit relation between constants. The empty set, and the empty relation, are represented by $\{\}$ and the universal set is represented by $\mathbf{Args}$. The universal relation is represented by ∞ and corresponds to the relation $\mathbf{Args} \times \mathbf{Args}$. Identity is added as the relation I, which obeys the

However, relation algebras do not have a similar correspondence to binary relations as defined as sets of pairs, and Boolean Modules likewise do not precisely capture the set theoretic interactions of sets and relations. As a result axiomatising generalised argumentation systems using Boolean Modules is sound but not complete.

It is possible to use extensions of Boolean Modules to axiomatise generalised argumentation systems. Pierce Algebras are the next natural candidate which extend Boolean Modules by one operation [22]. However Relation Algebras are inter-definability with Pierce Algebras [103] (essentially sets can be represented by special elements of the relation algebra), leading to a further sound but incomplete set of axioms.

Renate Schimdt studies the conditions under which Peirce algebras are *representable*, i.e. under which they behave like set theoretic binary relations [102]. The goal of an algebraic proof system would be to find suitable additional schemata or rules to augment Boolean Modules or Pierce algebras to provide a complete inference system.

[7] where the variables X_1 to X_m are the free variables defining solution sets

[8] Equally one could give any of the tableaux systems: tableaux, dual tableaux, signed tableaux; or give natural deduction rules.

For X and Y sets and J a relation

$$(X = Y)^T \mathrel{\widehat{=}} \forall x.((x \in X)^T \equiv (x \in Y)^T) \tag{7.1}$$

$$(X \subseteq Y)^T \mathrel{\widehat{=}} \forall x.((x \in X)^T \Rightarrow (x \in Y)^T) \tag{7.2}$$

$$(x \in \emptyset)^T \mathrel{\widehat{=}} \mathbf{false} \tag{7.3}$$

$$(x \in \mathbf{Args})^T \mathrel{\widehat{=}} \mathbf{true} \tag{7.4}$$

$$(x \in X \cup Y)^T \mathrel{\widehat{=}} (x \in X)^T \vee (x \in Y)^T \tag{7.5}$$

$$(x \in X \cap Y)^T \mathrel{\widehat{=}} (x \in X)^T \wedge (x \in Y)^T \tag{7.6}$$

$$(x \in \overline{X})^T \mathrel{\widehat{=}} \neg\, (x \in X)^T \tag{7.7}$$

$$(x \in J_{\exists}X)^T \mathrel{\widehat{=}} \exists y.((y,x) \in J)^T \wedge (y \in X)^T \tag{7.8}$$

$$(x \in J^{\forall}X)^T \mathrel{\widehat{=}} \forall y.((y,x) \in \breve{J})^T \Rightarrow (y \in X)^T \tag{7.9}$$

$$(x \in S)^T \mathrel{\widehat{=}} S(x) \text{ for } S \text{ a set variable} \tag{7.10}$$

$$\tag{7.11}$$

where $((y,x) \in J)^T$, for J a relation, is defined by:

$$((y,x) \in J_1 \cup J_2)^T \mathrel{\widehat{=}} ((y,x) \in J_1)^T \vee ((y,x) \in J_2)^T \tag{7.12}$$

$$((y,x) \in J_1 \cap J_2)^T \mathrel{\widehat{=}} ((y,x) \in J_1)^T \wedge ((y,x) \in J_2)^T \tag{7.13}$$

$$((y,x) \in \bar{J})^T \mathrel{\widehat{=}} \neg\, ((y,x) \in J) \tag{7.14}$$

$$((y,x) \in \breve{J})^T \mathrel{\widehat{=}} (x,y) \in J \tag{7.15}$$

$$((y,x) \in J)^T \mathrel{\widehat{=}} J(x,y) \text{ for } J \text{ a relation constant} \tag{7.16}$$

Figure 7.1.: First Order Translation

For relations J_1 and J_2

$$J_1 = J_2 \mathrel{\widehat{=}} \forall x,y.(x,y) \in J_1 \equiv (x,y) \in J_2 \tag{7.17}$$

$$(x \in \{a_1, a_2, \ldots, a_n\})^T \mathrel{\widehat{=}} x = a_1 \vee x = a_2 \ldots \vee x = a_n \tag{7.18}$$

$$\begin{aligned}((x,y) \in &\{a_1 \mapsto b_1, \ldots, a_n \ \mapsto b_n\})^T \mathrel{\widehat{=}} \\ &(x = a_1 \wedge y = b_1) \vee (x = a_2 \wedge y = b_2) \ldots \vee (x = a_n \wedge y = b_n)\end{aligned} \tag{7.19}$$

$$x \infty y \mathrel{\widehat{=}} \textbf{true} \tag{7.20}$$

Figure 7.2.: Constant Sets and Relations

usual conditions of reflexivity, symmetry, transitivity and substitution.

Thus we may specify a finite system by writing axioms of the form:

$$\begin{aligned}&\textbf{Args} = \{a_1, \ldots, a_n\}, \\ &\quad R_1 = \{a_1 \mapsto b_1, \ldots, a_m \ \mapsto b_m\}, R_2 = \{\ldots\}, \ldots R_k = \{\ldots\}, \\ &\quad S_1 = \{\ldots\}, \ldots, S_l = \{\ldots\}\end{aligned}$$

The additional translation rules are given by figure 7.2.

We adopt the standard structural rules for two-sided Gentzen systems which are given in figure 7.3

In writing the remaining Gentzen rules we will adopt the convention of leaving the context of the rule implicit. For example, if we take the left and right rules for disjunction, in which we normally write the context as the lists of formulas Γ and Δ, the rules that would be written with explicit context as:

$$\frac{\Gamma, \phi \vdash \Delta \quad \Gamma, \psi \vdash \Delta}{\Gamma, \phi \vee \psi \vdash \Delta} \vee\text{-left} \qquad\qquad \frac{\Gamma \vdash \phi, \psi, \Delta}{\Gamma \vdash \phi \vee \psi, \Delta} \vee\text{-right}$$

will be written instead as:

$$\frac{\phi \vdash \quad \psi \vdash}{\phi \vee \psi \vdash} \vee\text{-left} \qquad\qquad \frac{\vdash \phi, \psi}{\vdash \phi \vee \psi} \vee\text{-right}$$

The derived rules for the language of generalised argumentation systems are given in figures 7.3, 7.4, 7.5, 7.6 and 7.7. Throughout we use the convention that an undecorated variable (e.g. x) introduced "above the line" stands for any term and a decorated variable (e.g. x^*) stands for a new variable (also known as an arbitrary variable) that does not occur in any formula below the line.

Structural Rules:

$$\frac{\Gamma \vdash \Delta}{\Gamma, A \vdash \Delta}\ \text{Weaken-left} \qquad \frac{\Gamma \vdash \Delta}{\Gamma \vdash A, \Delta}\ \text{Weaken-right}$$

$$\frac{\Gamma, A, A \vdash \Delta}{\Gamma, A \vdash \Delta}\ \text{Contract-left} \qquad \frac{\Gamma \vdash A, A, \Delta}{\Gamma \vdash A, \Delta}\ \text{Contract-right}$$

$$\frac{\Gamma_1, B, \Gamma_2, A, \Gamma_3 \vdash \Delta}{\Gamma_1, A, \Gamma_2, B, \Gamma_3 \vdash \Delta}\ \text{Permute-left} \qquad \frac{\Gamma \vdash \Delta_1, B, \Delta_2, A, \Delta_3}{\Gamma \vdash \Delta_1, A, \Delta_2, B, \Delta_3}\ \text{Permute-right}$$

Figure 7.3.: Gentzen Structural Rules

This deduction system is a definitional extension of the usual system for first order predicate logic with equality (note: here equality on objects is formalised via the I-rules, the $=$-rules are really logical equivalence rules) and can either be used on its own or mixed with first order assertions and usual first order Gentzen rules. Considered as a system in its own right, the soundness and completeness of the proof system are inherited from the soundness and completeness of the Gentzen system for first order logic with equality[9]. Note that the sets and relations used here cannot be quantified over but that it is still possible to carry out schematic proofs i.e. proof for arbitrary sets and relations obeying some axioms.

Proofs of general properties about argumentation systems can be carried out in this formal system (for example, the proof of $R_\exists S \subseteq \overline{S}$ implies that $\breve{R}_\exists S \subseteq R_\exists S$); equally, given a particular finite argumentation systems, the formal system can be used to prove that a particular set is a solution set under a given semantics (for example, given that with **Args** $= \{a, b, c\}$ and $R = \{a \mapsto b, b \mapsto c\}$ then the set $S = \{a, b\}$ is conflict free and acceptable i.e. that $R_\exists S \subseteq \overline{S}$ and $\breve{R}_\exists S \subseteq R_\exists S$).

This answers the first question above. The second question, of what to do when minimisation/maximisation is used in the semantics raises the issue of quantifying over sets to express the relative "size" with respect to some ordering. This suggests that the first order translation, as used to derive the above rules, will be inadequate and that we must resort to a higher order logic, if we are to follow a similar approach. Two possibilities present themselves for investigation. The first is to see if the higher order aspect can be limited to monadic logic since we are only maximising (or minimising) sets, and these may be translated as unary predicates, with the ordering being implication. In this case one might derive additional sequent rules from weak second order monadic logic. Alternatively, if

[9] Here completeness means that if a property holds semantically then there is a proof of it. This is established by showing that if a backward proof search fails to terminate in an axiom because some branch of the backward search fails to close after all possible rules have been fairly applied then the resulting branch can be used to build a counter-model to the goal. This property is maintained provided that our proof rules extract all the "logical content" of the formula.

Top level relations:

$$\frac{x \in B \vdash \quad \vdash x \in A}{A \subseteq B \vdash} \subseteq_1\text{-left} \qquad \frac{x^* \in A \vdash x^* \in B}{\vdash A \subseteq B} \subseteq_1\text{-right}$$

$$\frac{xSy \vdash \quad \vdash xRy}{R \subseteq S \vdash} \subseteq_2\text{-left} \qquad \frac{x^*Ry^* \vdash x^*Sy^*}{\vdash R \subseteq S} \subseteq_2\text{-right}$$

$$\frac{x \in A, x \in B \vdash \quad \vdash x \in A, x \in B}{A = B \vdash} =_1\text{-left} \qquad \frac{x^* \in A \vdash x^* \in B \quad x^* \in B \vdash x^* \in A}{\vdash A = B} =_1\text{-right}$$

$$\frac{xRy, xSy \vdash \quad \vdash xRy, xSy}{R = S \vdash} =_2\text{-left} \qquad \frac{x^*Ry^* \vdash x^*Sy^* \quad x^*Sy^* \vdash x^*Ry^*}{\vdash R = S} =_2\text{-right}$$

Boolean Operations on Sets:

$$\frac{x \in A, x \in B \vdash}{x \in A \cap B \vdash} \cap\text{-left} \qquad \frac{\vdash x \in A \quad \vdash x \in B}{\vdash x \in A \cap B} \cap\text{-right}$$

$$\frac{x \in A \vdash \quad x \in B \vdash}{x \in A \cup B \vdash} \cup\text{-left} \qquad \frac{\vdash x \in A, x \in B}{\vdash x \in A \cup B} \cup\text{-right}$$

$$\frac{\vdash x \in A}{x \in \overline{A} \vdash} {}^{-}\text{-left} \qquad \frac{x \in A \vdash}{\vdash x \in \overline{A}} {}^{-}\text{-right}$$

Operators on Sets:

$$\frac{y^* \in A, y^*Rx \vdash}{x \in R_{\exists}A \vdash} {}_{\exists}\text{-left} \qquad \frac{\vdash y \in A \quad \vdash yRx}{\vdash x \in R_{\exists}A} {}_{\exists}\text{-right}$$

$$\frac{y \in A \vdash \quad \vdash xRy}{x \in R^{\forall}A \vdash} {}^{\forall}\text{-left} \qquad \frac{xRy^* \vdash y^* \in A}{\vdash x \in R^{\forall}A} {}^{\forall}\text{-right}$$

Figure 7.4.: Gentzen Rules for Sets

Boolean Operations for Relations:

$$\frac{xRy \vdash \quad xSy \vdash}{x\,(R \cup S)\,y \vdash}\ \cup\text{-left} \qquad\qquad \frac{\vdash xRy, xSy}{\vdash x\,(R \cup S)\,y}\ \cup\text{-right}$$

$$\frac{xRy, xSy \vdash}{x\,(R \cap S)\,y \vdash}\ \cap\text{-left} \qquad\qquad \frac{\vdash xRy \quad \vdash xSy}{\vdash x\,(R \cap S)\,y}\ \cap\text{-right}$$

$$\frac{\vdash xRy}{x\,\overline{R}\,y \vdash}\ {}^{-}\text{-left} \qquad\qquad \frac{xRy \quad \vdash}{\vdash x\,\overline{R}\,y}\ {}^{-}\text{-right}$$

Relational Algebra Operators on Relations:

$$\frac{\vdash yRx}{x\breve{R}y \vdash}\ \breve{}\text{-left} \qquad\qquad \frac{\vdash yRx}{\vdash x\breve{R}y}\ \breve{}\text{-right}$$

$$\frac{xRz^{*}, z^{*}Sy \vdash}{x\,(R \fatsemi S)\,y \vdash}\ ;\ \text{-left} \qquad\qquad \frac{\vdash xRz \quad \vdash zSy}{\vdash x\,(R \fatsemi S)\,y}\ ;\ \text{-right}$$

$$\frac{yRz^{*}, z^{*}Sx \vdash}{x\,(R \circ S)\,y \vdash}\ \circ\ \text{-left} \qquad\qquad \frac{\vdash yRz \quad \vdash zSx}{\vdash x\,(R \circ S)\,y}\ \circ\ \text{-right}$$

Identity Relation:

$$\frac{y\,I\,x \vdash}{x\,I\,y \vdash}\ \text{sym-left} \qquad\qquad \frac{\vdash y\,I\,x}{\vdash x\,I\,y}\ \text{sym-right}$$

$$\frac{y \in A \vdash}{x \in A, x\,I\,y \vdash}\ I_{\text{set}}\text{-left} \qquad\qquad \frac{\vdash y \in A}{x\,I\,y \vdash x \in A}\ I_{\text{set}}\text{-right}$$

$$\frac{zRt \vdash}{xRy, x\,I\,z \vdash}\ I_{\text{rel}}\text{-left}_{1} \qquad\qquad \frac{\vdash zRy}{x\,I\,z \vdash xRy}\ I_{\text{rel}}\text{-right}_{1}$$

$$\frac{xRz \vdash}{xRy, x\,I\,z \vdash}\ I_{\text{rel}}\text{-left}_{2} \qquad\qquad \frac{\vdash xRz}{x\,I\,z \vdash xRy}\ I_{\text{rel}}\text{-right}_{2}$$

Figure 7.5.: Gentzen Rules for Relations

Constant Sets and Relations:

$$\frac{x \in \{a_1\} \vdash \quad \ldots \quad x \in \{a_n\} \vdash}{x \in \{a_1, a_2, \ldots, a_n\} \vdash}\ \{\}\text{-left} \qquad \frac{\vdash x \in \{a_1\}, \ldots, x \in \{a_n\}}{\vdash x \in \{a_1, a_2, \ldots, a_n\}}\ \{\}\text{-right}$$

$$\frac{x\{a_1 \mapsto b_1\}y \vdash \quad \ldots \quad x\{a_n \mapsto b_n\}y \vdash}{x\,\{a_1 \mapsto b_1, \ldots, a_n \mapsto b_n\}\,y \vdash}\ \{\}\text{-left}$$

$$\frac{\vdash x\{a_1 \mapsto b_1\}y, \ldots, x\{a_n \mapsto b_n\}y}{\vdash x\,\{a_1 \mapsto b_1, \ldots, a_n \mapsto b_n\}\,y}\ \{\}\text{-right}$$

Figure 7.6.: Gentzen Rules for Constants

this is not possible then simple type theory (or a low order fragment of simple theory) might be adopted to define the maximisation/minimisation required.

Similarly, in the above logic, we can capture the arithmetic part of the axioms for reflexive transitive closures of relations i.e. for T the reflexive transitive closure of R, $I \subseteq T$, $R \subseteq T$ and $T \fatsemi T \subseteq T$, and indeed, could add the corresponding inference for an operator $\_^*$ on binary relations. This would be adequate for reasoning in a fixed finite system in the sense that for any pair of values connected by such a closure there would be a proof of that connection (since the proof would simply be the unwrapping of the closure into its steps). But the induction principle would be missing, i.e. that the closure is *the least relation* satisfying the axioms. Here again, the route of formalising in simple type theory and attempting to extract the logical content as rules seems a way forward.

Although both these areas need to addressed before we can arrive at a full deductive account of Trust Systems in general, the above proof system, with the arithmetic axioms for reflexive transitive closure, is sufficient for practical work on opinion consistent trust systems in the sense that it can be used to both prove the club conditions on a set, and the existence of a social connection within a set.

Extending the deductive logic to the Conjunctive Coalition Systems, introduced in Chapter 6, is of interest for at least two reasons. Firstly, intrinsically, because it extends argumentation systems from pair-wise interaction of augments to interaction between multiple arguments[10].

Secondly, from the results of Chapter 6 it seems probable that a Preferential

[10] Not only is such interaction interesting in its own right but it can also provide a simple account of higher order argumentation in which arguments may *attack an attack* rather than just another argument [77]. To see how this may be done consider a simpler system with a ternary attack relation $R : \mathbb{P}(\mathbf{Args} \times \mathbf{Args} \times \mathbf{Args})$, with the meaning of $R(a, b, c)$ being, taken together a and b attack c. We divide **Args** into two disjoint sets which we call *facts* and *reasons*. Facts play the role of arguments that may attack one another and may also attack reasons. Reasons are regarded as labels to attacks between facts and between facts and reasons. If we restrict R to $R : \mathbb{P}(\mathit{facts} \times \mathit{reasons} \times \mathbf{Args})$ then we have the skeleton framework for a higher order argumentation system.

Axioms:

$$\frac{}{P \vdash P}\ \text{axiom}$$

$$\frac{}{x \in \{\} \vdash}\ \{\}_{\text{set}}\text{-left}$$

$$\frac{}{\vdash x \in \mathbf{Args}}\ \mathbf{Args}\text{-right}$$

$$\frac{}{\vdash x \in \{x\}}\ \{-\}_{\text{set}}\text{-right}$$

$$\frac{}{x\,\{\}\,y \vdash}\ \{\}_{\text{rel}}\text{-left}$$

$$\frac{}{\vdash x \infty y}\ \infty\text{-right}$$

$$\frac{}{\vdash x\,I\,x}\ I\text{-right}$$

$$\frac{}{\vdash x\,\{x \mapsto y\}\,y}\ \{-\}_{\text{rel}}\text{-right}$$

Figure 7.7.: Gentzen Rules for Axioms

Coherence logic can be built from a logic for Conjunctive Coalition systems. Such an approach would make Preferential Coherence a special case of argumentation. Potentially the mix of Preferential Coherence and other argumentation would allow the construction of trust systems in which Knowledge-on-Trust and Social-Trust reasoning are combined.

The complicating factor here is that we need to quantify over sets of arguments, so that we may say, for example, that a coalition A attacks a coalition B if there is a subset of A that attacks a subset of B. Following the approach above, the appropriate method would seem to be to start with a framework of type theory, phrase a translation of the relational language into simple type theory, then create a system of derived rules that extract the logical content of the definitions.

Finally, an interesting direction to consider is proof systems for finite models i.e. what happens when we take models as finite but not specifically given. That is, can we get a proof system in which we may draw conclusions about what must be true for all finite sets of arguments when the relations have some particular properties?

7.4. Risk and Influence

Trust is taken *on-balance*, *all-things-being-equal*, on the basis of *nothing-to-the-contrary*. It is not absolute, it is not certain; rather it stands in place of certainty when certainty itself is absent. The attempt to find certainty is misplaced, a waste of effort that could be better spent by getting on with life. We are human, we co-operate, we trust. We start off trusting and learn to distrust, this is the fate of social animals. The analysis of the gains of trust are exactly that. An analysis, not an explanation of why we trust. Notions of reciprocation, and generalised reciprocation, cannot give rise to trust because before there is reciprocation, there must be a deferred return, a deferred part of an exchange, there must be a promise and there must be trust in that promise. We do not learn trust from reciprocation, rather we learn distrust from the failure of reciprocation. If we have gone so far in learning distrust that we do not trust strangers then how are we to deal with them? The social answer is to acquire trust from those who have unbroken trust. You are not a stranger and I trust you when you say this stranger is trustworthy. So, at least provisionally, I trust the stranger (that is all-things-being-equal, no-knowledge-to-the-contrary etc.). And in all this there is *risk*. The risk I am exposed to is that I was wrong to place faith in your opinion. So the question of risk is really the one of how much your opinion mattered in my decision to trust. Just how sensitive was my decision to your input?

In this section we consider a notion of the sensitivity of argumentation systems and trust systems. This analysis is based on that of Kenneth Parker and Edward McCluskey's analysis of probabilities for boolean switching functions, reported in "Probabilistic Treatment of General Combinational Networks" [86] and subsequent elaboration [70, 87]. Here, the work is extended to non-functional (that is, relational) systems[11]. The analysis below develops the notion of the degree of

[11] It should perhaps be pointed out that variations on this functional analysis have been independently discovered many times in areas such as, voting analysis, in which it has been used to

a	b	$c = a \wedge b$
1	1	1
1	0	0
0	1	0
0	0	0

Figure 7.8.: Truth Table for And

influence that one argument, or individual, being in a position, has upon another argument, or individual, being in a position.

In an argumentation system, arguments which are not attacked or supported by any other argument are *independent variables* whose values may be freely set to true or false. An argument which is not an independent variable is called a *dependent variable*. Dependent variables may not be wholly determined by the independent variables but are constrained by them[12]. In trying to formulate a position for an argumentation system we might consider asking how much influence a particular independent variable has on a particular dependent variable being in a position. Extending this to trust systems amounts to asking how much influence a particular independent variable (an individual) has on a trust outcome. A complication that arises when considering trust systems is that the independent variable may affect the social connections between two individuals by affecting trust in any individual in the social connections. So in the case of a trust system we ask the question as to how much effect an independent variable has on a particular social connection between individuals.

First the case of influence in argumentation systems. We consider generalised argumentation systems via their boolean network representation.

If we consider a boolean network that represents a function from some inputs to some output and ask the question of how much influence a particular input has on the output, the answer is reasonably straightforward. The standard approach to this problem is to calculate in what percentage of cases changing only the chosen input causes the output to change. That is, if f is an n-argument function, then the influence of the i^{th} variable x_i is:

$$\frac{\left|\left\{ (x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n) \,\middle|\, \begin{array}{l} f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n) \neq \\ f(x_1,\ldots,x_{i-1},\overline{x_i},x_{i+1},\ldots,x_n) \end{array} \right\}\right|}{2^n}$$

As a trivial illustration consider $c = a \wedge b$:

Examining the truth table in figure 7.8 we see that if $b = 1$ then if a changes

determined how much influence a single vote carries in a voting scheme and how that influence scales with the population size. Possibly the earliest use of this is by L. S. Penrose in 1946 [88] to derive that the influence of a vote in a simple majority choice scales as $\sqrt{n}$.

[12] We should observe at this point that taking the arguments that are neither attacked nor supported by other arguments is a *convention*. In fact we are free to take any set of arguments as "independent" and regard their value assignments as constraining the rest of the argumentation network. However, given the intended meaning of attack and support the convention seems worthwhile.

a	b	$a \Rightarrow b$
1	1	1
1	0	0 ***
0	1	1
0	0	1

Figure 7.9.: Truth table for Implication Relation

from 1 to 0 then c must change from 1 to 0 and if a changes from 0 to 1 then c must change from 0 to 1. In all other cases changing a does not require a change to c. a's influence on c is the number of rows of the truth table in which changing a requires a change in c, which is 2, divided by the number of rows, which is 4, so a's influence on c is $\frac{1}{2}$.

For argumentation systems the relationship between independent variables and dependent variables is not necessarily functional. As a result the "truth table" connecting the independent variables to a dependent variable becomes relational, i.e. it may contain two rows with the independent variables taking on the same values for the different values of the dependent variable, and some values of the independent variable may have no assignment to dependent variables.

To illustrate this let us consider the relation $a \Rightarrow b$. That is, the implication is required to be true so the truth table connecting a and b is shown in figure 7.9.

As we require the implication to hold, the starred row is banned since it represents a case when the implication is false. As a result when $b = 0$ and a changes from 0 to 1, b must change from 0 to 1 if the implication is to be maintained (and this is the only case where changing a whilst maintaining the relation of implication between a and b requires b to change). The influence of a on b is therefore $\frac{1}{3}$. The general case is that we have an axiomatically constrained boolean theory in a set of independent variables $(x_1, \ldots, x_n)$ and a dependent variable d. This theory has a set of models M which we may write as vectors $(d, x_1, \ldots, x_n)$. The influence of an independent variable x_i on the dependent variable d is:

$$\frac{\left| \left\{ (d, x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) \middle| \begin{array}{l} (d, x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) \in M \wedge \\ (d, x_1, \ldots, x_{i-1}, \overline{x_i}, x_{i+1}, \ldots, x_n) \notin M \end{array} \right\} \right|}{\mid M \mid}$$

However, for argumentation this is not the whole story. We may be interested in the risk associated with the assumption that the independent variable is assumed to have a particular value. For example, if we assume an independent variable is true what is the likelihood of this leading to including some dependent variable in error? That is, what is the likelihood that changing the assumption will change the inclusion of the dependent variable. Let us adopt the names influence$_0$ and influence$_1$ for the influence of an independent variable on a dependent variable in changing from 0 to 1 and analogously for the change from 1 to 0. As illustration, in the implication example of figure 7.9, influence$_0 = \frac{1}{3}$ and influence$_1 = 0$. Clearly, infulence$_0$ is given by:

$$\frac{\left|\left\{(d,x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)\;\middle|\;\begin{array}{l}(d,x_1,\ldots,x_{i-1},0,x_{i+1},\ldots,x_n)\in M\wedge\\(d,x_1,\ldots,x_{i-1},1,x_{i+1},\ldots,x_n)\notin M\end{array}\right\}\right|}{\mid M\mid}$$

and influence$_1$ by:

$$\frac{\left|\left\{(d,x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)\;\middle|\;\begin{array}{l}(d,x_1,\ldots,x_{i-1},1,x_{i+1},\ldots,x_n)\in M\wedge\\(d,x_1,\ldots,x_{i-1},0,x_{i+1},\ldots,x_n)\notin M\end{array}\right\}\right|}{\mid M\mid}$$

and the total influence is the sum of influence$_0$ + influence$_1$. If the influence of an independent variable on a dependent variable is 1 then the two variables are completely correlated and the independent variable acts as a switch controlling the dependent variable.

To apply the notion of influence to paths in trust systems we note that the existence of a path can be modelled by a new dependent variable whose value is the conjunction of the variables in the path. Applying this idea to the CA examples in Chapter 5 we see that the variable News controls the trust in the company websites of the examples.

This form of sensitivity analysis offers a method of assessing the risk we take in making assumptions about the trustworthiness of individuals that are not otherwise constrained by the network of relationships.

One intriguing possibility for further research is to use the notion of influence to generate approximations to argumentation systems in general and trust systems in particular. That is, we may generate simplified networks in which the influence of selected independent variables on selected dependent variables is approximately the same as the original system but lacking a number of the unselected variables.

Note: Moreover, the more advanced technique, spectral analysis of boolean functions (see, for example, O'Donnell's "Some topics in analysis of boolean functions " [80]) permit one to define further measures on functions such as:

Bias: the degree to which a particular outcome is favoured;

Energy: The average "swing input" i.e on average how many inputs need to change to change the outcome;

Noise Stability: the probability that the output would remain the same even if there was an independent input error probability of e for each input.

Extending these notions to relations offers the possibility of a more detailed analysis of the behaviour of particular argumentation and trust systems.

7.5. Formal Logic of Promises

The theory of promises was informally developed in Chapter 2 as a basis for discussing trust. The theory involves the interaction of actors via the promises made between them, the belief states of actors and the choices and actions of the actor, including choosing to trust and the fulfilment, or otherwise, of the promises. A formalisation of a logic of this system would have much in common with logics for protocols, concurrency and, so called, agent logics, since they must discuss similar notions, e.g. communication events, internal states of belief, actions, etc. However the logic must differ in certain areas. It must encompass rational belief based on uncertain knowledge and so non-monotonicity and/or paraconsistency; it must encompass commitment as well as communication to capture consequential promises (i.e. promises that have consequences if they are not fulfilled); and it must encompass the choice to trust, i.e. the distinction between the potential trustee simply performing an action from which the potential truster simply takes some advantage and the action of the trustee being *relied upon* by a truster. The main technical question to be addressed, however, is to what degree one can make promises about promises and how to resolve the potential self-reference if this is allowed. The logic of promises also takes us towards a logic of anticipatory systems in that current actions may be based on the expectation, i.e. the anticipation, of the fulfilment of promises by another party. To understand the significance of this for trust consider the centipede game of Chapter 2. The backward induction is avoided by each player anticipating the co-operation of the other. But this anticipation is itself underwritten by trust. So we see may expect *anticipatory logic*, and the full logic *trust* and *promises*, to be intertwined in their development.

Given such a logic we might move from the trust interaction diagrams given in Chapter 2 to trust models which explain the dynamics of different forms of trust and formalise by what rational basis we can avoid *the collapse of trust*, that so often arises in the game theoretic analysis of interactions.

7.6. The Dynamics of Trust

This dissertation has been concerned with finding stable solutions to the problem of Knowledge-on-Trust and Social-Trust. But what of the dynamics of how these solutions are arrived at? If we consider, for example, Social-Trust, one might ask how a stable trust assignment might evolve from a random assignment. In general, such a process need not find a fixed point and may lock onto a cycle. The study of the dynamics of social trust evaluation may illuminate pathologies of trust in the real world, e.g. situations in which trust may fleetingly appear, only to disappear and never be seen again. Such situations may arise because "local conditions" initially give rise to trust but distant conditions overturn this local outcome. And equally there may be insights into the opposite situation, in which distrust in an individual gives way to trust as increasing social knowledge is obtained from distant parts of one's social network. Mathematically this problem seems closely connected to the problem of solving satisfiability problems by finding fixed points in a boolean network [23, 94].

If one shifts from an individual to a group perspective then one may further imagine the dynamics as each individual re-evaluates and updates their position in relation to the updates of other individuals. Which then itself leads to further questions of how synchronous update versus asynchronous update differ in outcome. The two processes are analogous to different social processes. For example, revision and update as the result of a news broadcast causes many individuals to more-or-less simultaneously change their opinions, whereas update driven by person-to-person spread rumour is essentially asynchronous. Clearly the dynamics of trust offers many possible avenues to explore in future work.

7.7. The Elephant in the Room

As this research has continued it has become clear that there are many overlaps with network theories in the social sciences. These connections did not show up in my initial researches into theories of trust and trustworthiness but rather emerged as I investigated a mathematical and computational theory of social trust. This begs two questions: Why did such connections not emerge earlier? What, if anything, is different about the theory developed here and other theories of social networks?

The answer to the first question seems to be cultural. The majority of social scientists and philosophers have tended to view trust informally and as a phenomenon of individual psychology, with social context playing only a small role. Those that have explored formal models have employed probability models based on the notion that to trust is to take a risk, or have adopted game theoretic models to explain the gains of trust versus distrust (i.e. some form of reciprocity model). In these approaches the analysis is either in terms of pairwise interactions, or in terms of a generalised group, or category, of interactants.

In contrast the network theorist have produced formal models in which pairwise interactions form a network but the pairwise interactions they have considered are the more-or-less objective properties of social groups, such, as connectivity, influence, alliance, opposition, etc. In these models rather than considering generalised categories of interactants, individuals interact with individuals to whom they are explicitly connected to by some defined relation. Trust does not seem to be a relation of interest to social network theorists, although one may argue it is the key underlying relation assumed in the notions of alliance, friendship and kinship (i.e., one assumes the significance of, say, a bond of friendship, in that it entails that one participant trusts the other with respect to certain actions).

Rarely do these styles of analysis meet and where they do (see below) trust is not explicitly discussed. In particular, in answer to our first question, no-one, it seems, has attempted to answer questions of trust through network theory.

Where the styles of analysis do meet is in Harrison White's notion of catnets [114], and Dorwin Cartwright and Frank Harary's Structural Balance Theory [25].

A catnet is the bringing together of "the notions of network (abridged net) and category (abridged cat) in a new concept, catnet, which can be defined as any set of individuals comprising both a category and a network" (Santoro [99]) or "a bunch of people alike in some respect, from someone's point of view" (White [114]) who are connected by a relationship between pairs of individuals where

i) "the persons in the net accept the idea that they have meaningful indirect relations with anyone paired with a man with whom they are paired (i.e., most of all persons, when considered as the ego viewing the net, accept "composition" as a meaningful "operator" on the net, to use mathematical language);

ii) the relation diagrammed implies sufficient familiarity with the other in a pair to have some idea of who else he is related to;

iii) yet the many possible indirect connections among people, at various removes (through various numbers and patterns of intermediary persons) are not recognised as falling into distinct new types of relations with their own definitions and contents." (White [114]).

So that, for example, we may consider a category, say neighbours, and a network relation, say friendship, to determine a catnet of socialisation and co-operation in my neighbourhood. Here the analysis differs from, say, that based on notions such as solidarity (based for example on common, or shared, identity), in that it is not simply a common property that determines the scope of socialisation and co-operation but the particular network structure that exists between individuals i.e. the particular structure of connections is important to understanding what can, and cannot happen.

The second notion, Structural Balance Theory, is a theory of stable social interaction that says social relationships are formed in such a way as to minimise a certain form of stress in relationships. In essence it says that we form "friends" and "enemies" and we attempt to maintain the triangle rules that "the friend of my friend is my friend" and "the enemy of my enemy is my friend". The theory was introduced and formalised using signed graph theory by Dorwin Cartwright and Frank Harary in 1956 [25], and since then has been studied extensively (see, for example, Frank Harary and Robert Norman's "Structural Models: An introduction to the theory of Directed Graphs" [47] and Per Hage and Frank Harary "Structural Models in Anthropology" [45], or Chapter 6 of Stanley Wasserman and Katherine Faust's "Social Network Analysis" [113]).

Signed graphs are graphs with edges labelled $+$ or $-$ (or $+1$ and -1). Paths are signed by multiplying the signs of the edges in the path. A half-edge is an edge with a single vertex. A signed graph is called balanced if it has no half-edges and every cycle has a positive sign. The central theorem of signed graph theory is that a signed graph is balanced if, and only if, its vertices can be divided into two sets, either of which may be empty, such that all edges within a set are positive and all edges between the sets are negative. Signed graphs and the balance theorem are used in social modelling as Structural Balance Theory to model the interaction of groups with positive and negative connections. Balanced graphs are viewed as stable and unbalanced graphs as unstable.

A weaker form of structural balance has also been posited by James Davis [30] which only requires the first triangle rule, "the friend of my friend is my friend", thus permitting the possibility that "the enemy of my enemy is my enemy". The equivalent classification theorem becoming "if a signed graph is weakly structurally balanced then it can be divided into groups in such a way that all edges between members of a group are positive and all edges between members of different groups are negative.

Each of these notions has similarities to the approach taken to Social Trust based on argumentation theory. In each case the notion of network is central, so that the range of possibilities of individual action, or beliefs, arises from their position in a network rather than from being part of some amorphous group or class. In Structural Balance Theory, the stable sets that arise are the result of a resolution of positive and negative forces on individuals, and the resolution can be seen as a striving for consistent sets.

There are also differences. Social Trust theory, in contrast with catnets, lacks a notion of category. Mathematically, this lack of the notion of category has no great significance. After all, one could simply introduce the notion of categories via an equivalence relation on individuals. Two individuals are equivalent if, and only if, they are in the same category. Then, for example, trust might be restricted to equivalent individuals and distrust is automatic between non-equivalent individuals. That is, only individuals in the same category can be trusted. The more important question is should the social theory include a notion of category? Should category membership play a role in establishing the grounds for trust and distrust? The role of category is to establish a pre-condition beyond the direct or transitive connections of the network. In this sense it may capture a notion such as class solidarity. But, in the framework of the social trust theory it would be a bias and not an absolute i.e. we would trust unless there was a reason to distrust. In which case, however, there is no real difference between this notion and the implicit trust discussed previously. That is, it is simply the complement of the distrust relation. One view of this is to say that we have just restated the mathematical equivalence. Another view, however, is that social category can also be thought of via relationships between individuals rather than as a separate concept. That is, social categories themselves arise out of the relationships between individuals.

Similarly, Social Trust theory, when compared with Structural Balance Theory, recognises the existence of competing 'forces' on individuals but does not assume an evolution to a resolved situation. Rather it accepts that individuals may be members of many different consistent communities, each of which can constitute an alternative basis for a relation of initial trust in an individual via a social connection within that community. Moreover the actual definition of consistency differs between the various notions of Structural Balance and that of Social Trust.

Answering the second of the two questions then: The theory developed herein is different from what appear to be the most relevant network theories. It makes different social assumptions and different mathematical assumptions.

This said, it is undoubtedly true that a thorough investigation of the links and differences between the notion of social trust and more mainstream social network theories would be of considerable interest, as would an investigation of the connection with the notion of Relational Sociology (see for example, Mustafa Emirbayer's "Manifesto for a Relational Sociology" [35]). Alas, this must be left to future research.

7.8. Conclusion

It is hoped that the topics covered in this chapter have gone some way to illustrating both the breadth and depth to be found in the logical modelling of trust and trustworthiness. The modelling approach has been one of unravelling logical relationships between individuals or between utterance of individuals. On the theoretical side, this gives rise to questions of how one reasons about these relationships. When numeric considerations arise, they arise as intrinsic measures of properties of the relationships, rather than as numeric indicators of uncertainty. On the practical side, focusing on relationships forces us to consider the possible sources of such information and the problems of individuals disclosing such information for practical ends.

Thematically, the link to mathematical sociology offers many possibilities for the future, not merely because of the overtly computational nature of the network view of social interaction, but even more promisingly, because of the potential applicability of the tools of theoretical computer science to the study of these network structures.

Clearly, however, much must await future work. In the final chapter we will turn to what has been gained from the journey so far.

8. In Conclusion

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.

Little Gidding V

T. S. Eliot

In which is discussed:

- *Where we have arrived.*
- *What has been gained from the journey.*

8.1. To Recap

Previous chapters have introduced a view of trust as arising from social embeddedness that makes it distinct from issues of security. Technical infrastructure can offer security but not trust or trustworthiness. Trust is the decision to act on a belief in the trustworthiness of an individual. Specifically it is the decision to act on the basis that the promises made by the trusted individual will be fulfilled.

We have three sources of gaining information about trustworthiness:

- Our own direct experience;
- The experience of others as reported to us;
- The reports of the trustworthiness of others as reported to us.

Our problem in the modern world is to bootstrap from our own direct experience, using information gathered via our social network, to establish the trustworthiness, or otherwise, of people of whom we have no (or insufficient) direct experience.

In this dissertation I have posed two, related, sub-problems of bootstrapping trust: How can one obtain Knowledge on Trust? And, How can one establish Social Trust? These questions have been answered by developing both informal and formal analysis of the concepts inherent in the questions and developing effective, i.e. computational, means of performing the bootstrapping. The computational approach has been based on adapting standard reasoning tools for classical logic (i.e. model checkers and proof systems in the form of tableaux) to the new reasoning problem.

Throughout the dissertation I have deliberately emphasised those aspects of trust arising through social processes as opposed to cognitive processes because, in a modern society, trust rarely arises out of direct knowledge of other individuals. This emphasis on social embeddedness is new and distinct from both the *traditional* approach to trust which emphasises the experience of an individual, and from *swift trust* which emphasises an individual's experience of roles.

But where does all this get us? That is question to which we now turn in the final section.

8.2. What is the Significance of this Theory of Trust?

Why is the theory of trust developed here significant? What might we do differently in the light of such a theory?

There are two forms of answer that we dub the "immediately practical" and the "contextual" answers.

Of the "immediately practical", two answers are foremost:

- Because we want to improve the safety of remote interaction via the Internet and the web;
- Because computer access control models are ultimately based upon, and built around, an understanding of trust.

The first of these has been discussed in section 7.2 above. The latter answer needs some elaboration, for it may not be obvious. The claim here is not that the developers of access control models explicitly took a particular model of trust and "implemented it", but rather that particular models of trust, implicitly or explicitly, underlie their decisions, and that these correspond to models of trust developed elsewhere.

The original Bell-Lapadula model of confidentiality [7, 8, 9] was built on a notion of explicit trust in which a vetting authority has direct trust in individuals to handle documents at certain classifications. The Brewer-Nash, Chinese-Wall, model of access control [20] kept with this explicit trust model but added dynamics to the state of an individual (i.e. whether the individual is in possession of/has accessed particular information). Ferraiolo and Kuhn's Role Based Access Control [39], is a model of access control based on Swift Trust, i.e. it is not the individual but the role they occupy that is trusted to access classified information. Systems trust in the role assignment mechanism underwrites this access, in this case meaning that the vetting authority is trusted to correctly assign role entry permissions. Extended versions of RBAC include delegation that allows others than a centralised vetting authority to delegate or transfer access rights to others and so more closely correspond to the flow of trust in social structures and organisations.

If then we develop alternative models for trust and trustworthiness, that apply in new contexts or analyse trust and trustworthiness as arising in new ways, then we may expect there to be correspondingly new access control models. In particular we might consider new access control models built on either Knowledge-on-Trust or Social-Trust. Indeed, we may see the privacy model outlined in section 5.14 as exactly such a model built on Social-Trust.

Such are the immediate practical answers. The contextual answers are more indirect in nature. By the “contextual” answer I mean that which alters the context in which we do things and thereby alters what we do. By studying trust, and examining new models of trust acquisition, we alter our perspective on systems analysis. By considering trust as, at least in part, arising from our social embeddedness, we change our perspective on trust, from one of individual rationality, to one of social rationality in which we consider how an individual sits in our social world.

For example, in buying an item from a supplier that does not have the item in stock, I have the promise of the supplier that in return for my payment, *made now*, that the supplier will deliver the goods at some bounded time *in the future*. But I also have a promise from “civil society” to the effect that if the supplier does not deliver the goods, then there are actions that can be taken against the supplier. However, to secure the benefits of this second promise, I need proof of the transaction, and therefore a secure record of the transaction, including the jurisdiction it took place in, agreements entered into, etc. We take all this for granted because buying something from a supplier is commonplace. But the framework of good practice, social norms, case law and legislation governing such simple transactions and underwriting trust in the transaction, has taken long to establish and varies from jurisdiction to jurisdiction.

Yet, almost daily, we are responsible for creating new mechanisms of interaction and trade between people. The newness may arise because of relative anonymity, the size of audience reached, because of trade across jurisdictional borders, because it deals with new forms of content or commodity (e.g. information commodities with no tangible form), because new payment models are involved or any number of other forms of newness. Along with these mechanisms come new social conventions, new business opportunities, and new crimes. And these give rise, in their turn, to new laws, new responsibilities for individuals and new realms of regulation by governments. The changes we make are not trivial. They alter the dynamics of our society, often in unexpected ways[1].

Whatever the newness involved, it is the job of the systems analyst[2] to identify what promises are made by whom, and to whom they are made; what needs to be recorded and protected; and what needs to be audited and by whom, so that trust is appropriately underwritten.

That is, systems analysis is not performed in a vacuum. Rather it is performed against theories that fill in the “invisible” parts of the system. Having a different theory of trust and trustworthiness means we perform our analysis differently. As

[1] As may be illustrated by two minor examples of the unexpected behaviour. Danah Boyd’s doctoral dissertation gives an account of how account passwords for social network sites are shared as a token of intimacy [19] and the BBC news reports on how iTunes passwords are left in Wills to pass on music collections as part of an inheritance now that physical media are disappearing *BBC News 14 Oct 2011, http://www.bbc.co.uk/news/technology-15292748*. Neither of these possibilities are foreseen in the terms and conditions of websites or digital retailers.

[2] An old fashioned term, but I know of no other that appropriately captures the role.

Frederic Bastiat says of economists

> *There is only one difference between a bad economist and a good one: the bad economist confines himself to the visible effect; the good economist takes into account both the effect that can be seen and those effects that must be foreseen*[3].

So it is with systems analysts.

And it is for this, if for no other reason, that we must study trust. So that we may create theories of trust that are closer to trust-in-the-wild, so that we may be good systems analysts and foresee that which is not visible.

[3] Bastiat, Frederic. "What Is Seen and What Is Not Seen." Online at: http://www.econlib.org/library/Bastiat/basEss1.html.

Appendices

A. Proofs

PROPOSITION 1

- Acceptability - defined as $\breve{R}_\exists S \subseteq R_\exists S$ is equivalently expressed as $S \subseteq \breve{R}^\forall R_\exists S$.

by $\breve{R}_\exists S \subseteq R_\exists S$

hint Galois

$\equiv\ S \subseteq \breve{R}^\forall R_\exists S$

- Completness - defined as $D(S) \subseteq S$, is equivalently expressed as $\breve{R}^\forall R_\exists S \subseteq S$.

by $D(S) \subseteq S$

$\equiv\ \{y \mid \breve{R}_\exists \{y\} \subseteq R_\exists S\} \subseteq S$

$\equiv\ \{y \mid \{y\} \subseteq \breve{R}^\forall R_\exists S\} \subseteq S$

$\equiv\ \forall y.\, \{y\} \subseteq \breve{R}^\forall R_\exists S \Rightarrow y \in S$

$\equiv\ \forall y.\, y \in \breve{R}^\forall R_\exists S \Rightarrow y \in S$

$\equiv\ \breve{R}^\forall R_\exists S \subseteq S$

PROPOSITION 2 S is conflict free is equivalent to $\breve{R}_\exists S \subseteq \overline{S}$

by $R_\exists S \subseteq \overline{S}$

$\equiv\ S \subseteq \overline{R_\exists S}$

$\equiv\ S \subseteq \breve{R}^\forall \overline{S}$

$\equiv\ \breve{R}_\exists S \subseteq \overline{S}$

PROPOSITION 3 $S \cup R_\exists S = \textbf{Args}$ is equivalent to $\overline{S} \subseteq R_\exists S$.

by $S \cup R_\exists S = \textbf{Args}$

$\equiv\ \overline{S \cup R_\exists S} = \overline{\mathbf{Args}}$

$\equiv\ \overline{S} \cap \overline{R_\exists S} = \emptyset$

$\equiv\ \overline{S} \subseteq R_\exists S$

PROPOSITION 4 If S is admissible and $S \cup R_\exists S = \mathbf{Args}$ then it is preferred and semi-stable.

S is preferred follows from all greater S' are not admissible.

note from conflict free and $S \cup R_\exists S = \mathbf{Args}$ we have $\overline{S} = R_\exists S$

assume $S' \supset S$

hence $S' = S \cup S''$ for some $S'' \subseteq R_\exists S$

hence S' not conflict free and therefore not admissible.

and since S is preferred and $S \cup R_\exists S = \mathbf{Args}$ there can be no larger set $T \supset S$ in the coverage ordering.

PROPOSITION 5 If S is conflict free and stable then it is acceptable and complete.

S is acceptable

by $\breve{R}_\exists S \subseteq \overline{S}$

$\equiv\ S \subseteq \breve{R}^\forall \overline{S}$

hint $R_\exists S = \overline{S}$

$\equiv\ S \subseteq \breve{R}^\forall R_\exists S$

S is complete

by $\overline{S} = R_\exists S$

$\Rightarrow\ \overline{S} \subseteq R_\exists S$

$\equiv\ \overline{R_\exists S} \subseteq S$

$\equiv\ \breve{R}^\forall \overline{S} \subseteq S$

hint $R_\exists S = \overline{S}$

$\equiv \breve{R}^\forall R_\exists S \subseteq S$

PROPOSITION 7 $X \subseteq T^*_\exists X$

by $X \subseteq T^*_\exists X$

$\equiv X \subseteq X \cup T_\exists X \cup T^2_\exists X \cup \ldots$

$\equiv$ **true**

and hence, if X is trust closed, $X = T^*_\exists X$ by

by $X \subseteq T^*_\exists X$

$\wedge\ T^*_\exists X \subseteq X$

$\equiv X = T^*_\exists X$

PROPOSITION 8 X is internally trust consistent is equivalent to: the attackers of X are in the complement of the trust closure of X i.e. $\breve{A}_\exists X \subseteq \overline{T^*_\exists X}$.

by $(A \circ T^*)_\exists X \subseteq \overline{X}$

$\equiv A_\exists T^*_\exists X \subseteq \overline{X}$

$\equiv T^*_\exists X \subseteq A^\forall \overline{X}$

$\equiv T^*_\exists X \subseteq \overline{\breve{A}_\exists X}$

$\equiv \breve{A}_\exists X \subseteq \overline{T^*_\exists X}$

PROPOSITION 9 X is externally trust consistent is equivalent to: the attackers of the trust closure of X are in the complement of the trust closure i.e. $\breve{A}_\exists(T^*_\exists X) \subseteq \overline{T^*_\exists X}$.

by $(A \circ T^*)_\exists X \subseteq \overline{T^*_\exists X}$

$\equiv\ A_\exists T^*_\exists X \subseteq \overline{T^*_\exists X}$

$\equiv\ T^*_\exists X \subseteq A^\forall \overline{T^*_\exists X}$

$\equiv\ T^*_\exists X \subseteq \overline{\breve{A}_\exists T^*_\exists X}$

$\equiv\ \breve{A}_\exists (T^*_\exists X) \subseteq \overline{T^*_\exists X}$

PROPOSITION 10 A set X is externally trust consistent if its trust closure is conflict free, i.e. $A_\exists(T^*_\exists X) \subseteq \overline{T^*_\exists X}$.

by $(A \circ T^*)_\exists X \subseteq \overline{T^*_\exists X}$

$\equiv\ A_\exists T^*_\exists X \subseteq \overline{T^*_\exists X}$

PROPOSITION 11 If a set is externally trust consistent then it is internally trust consistent.

by $(A \circ T^*)_\exists X \subseteq \overline{T^*_\exists X}$

$\equiv\ (A \circ T^*)_\exists X \subseteq \overline{X \cup T_\exists X \cup T^2_\exists X \cup \ldots}$

$\equiv\ (A \circ T^*)_\exists X \subseteq \overline{X} \cap \overline{T_\exists X \cup T^2_\exists X \cup \ldots}$

$\Rightarrow\ (A \circ T^*)_\exists X \subseteq \overline{X}$

PROPOSITION 12 If a set is internally trust consistent then it is conflict free.

by $(A \circ T^*)_\exists X \subseteq \overline{X}$

$\equiv\ A_\exists (T^*_\exists X) \subseteq \overline{X}$

$\equiv\ A_\exists ((X \cup T_\exists X \cup T^2_\exists X \cup \ldots)) \subseteq \overline{X}$

$\equiv\ (A_\exists X \cup A_\exists T_\exists X \cup A_\exists T_\exists^2 X \cup \ldots) \subseteq \overline{T_\exists^* X}$

$\Rightarrow\ A_\exists X \subseteq \overline{X}$

PROPOSITION 13 If X is trust closed and internally trust consistent then it is externally trust consistent.

by $T_\exists^* X = X$

and $(A \circ T^*)_\exists X \subseteq \overline{X}$

$\Rightarrow\ (A \circ T^*)_\exists X \subseteq \overline{T_\exists^* X}$

PROPOSITION 14 X is externally trust consistent is equivalent to:

A: $X \subseteq {T^*}^{\forall} \breve{A}^{\forall} \circ {\breve{T^*}}^{\forall} \overline{X}$ and hence to

B: $\breve{T^*}_\exists \breve{A}_\exists {T^*}_\exists X \subseteq \overline{X}$ and

C: ${T^*}_\exists A_\exists \breve{T^*}_\exists X \subseteq \overline{X}$

A by $(A \circ T^*)_\exists X \subseteq \overline{T_\exists^* X}$

$\equiv\ T_\exists^* X \subseteq \overline{(A \circ T^*)_\exists X}$

$\equiv\ T_\exists^* X \subseteq (A \,\breve{\circ}\, T^*)^{\forall} \overline{X}$

$\equiv\ T_\exists^* X \subseteq (\breve{T^*} \circ \breve{A})^{\forall} \overline{X}$

$\equiv\ X \subseteq {T^*}^{\forall} (\breve{T^*} \circ \breve{A})^{\forall} \overline{X}$

hint $(U \circ V)^{\forall} = V^{\forall} \circ U^{\forall}$

$\equiv\ X \subseteq {T^*}^{\forall} \breve{A}^{\forall} \circ {\breve{T^*}}^{\forall} \overline{X}$

$\equiv\ X \subseteq {T^*}^{\forall} \breve{A}^{\forall} {\breve{T^*}}^{\forall} \overline{X}$

B by $X \subseteq {T^*}^{\forall} \breve{A}^{\forall} {\breve{T^*}}^{\forall} \overline{X}$

hint Galois

$\equiv\ \breve{T}^{*}{}_{\exists}\breve{A}_{\exists}T^{*}{}_{\exists}X \subseteq \overline{X}$

C by $X \subseteq T^{*\,\forall}\breve{A}^{\forall}\breve{T}^{*\,\forall}\overline{X}$

hint conjugates

$\equiv\ \breve{T}^{*}{}_{\exists}A_{\exists}T^{*}{}_{\exists}X \subseteq \overline{X}$

PROPOSITION 15 If X is a Trust Extended admissible set then it is a subset of some Y which is a Trust Closed admissible set, i.e.

$$\textbf{admissible}_X\, X \Rightarrow \exists\, Y \supseteq X.\, \textbf{admissible}_T\, Y.$$

by $\textbf{admissible}_X\, X \Rightarrow \exists\, Y \supseteq X.\, \textbf{admissible}_T\, Y.$

hint Choose $Y = T^{*}_{\exists}X$ and simplify

$\Leftarrow\ \textbf{admissible}_X\, X \Rightarrow T^{*}_{\exists}T^{*}_{\exists}X \subseteq T^{*}_{\exists}X \wedge \textbf{admissible}_D\, T^{*}_{\exists}X$

hint $T^{*}_{\exists}$ is a closure operator

$\equiv\ \textbf{admissible}_X\, X \Rightarrow \textbf{admissible}_D\, T^{*}_{\exists}X$

$\equiv$ **true**

PROPOSITION 17 X is Externally Consistent is equivalent to there exists some Y subset of X such that Y is T-closed and conflict free.

$\Rightarrow$ **forward direction**

by $A_{\exists}T^{*}_{\exists}X \subseteq \overline{T^{*}_{\exists}X}$

hint $T^{*}_{\exists}T^{*}_{\exists}X \subseteq T^{*}_{\exists}X =$ **true**

$\equiv\ T^{*}_{\exists}T^{*}_{\exists}X \subseteq T^{*}_{\exists}X \wedge A_{\exists}T^{*}_{\exists}X \subseteq \overline{T^{*}_{\exists}X}$

hint abstract over $T^{*}_{\exists}X$

$\Rightarrow\ \exists\, Y. T^{*}_{\exists}Y \subseteq Y \wedge A_{\exists}Y \subseteq \overline{Y}$

$\Leftarrow$ **backward direction**

by $\exists Y . T_{\exists}^{*} Y \subseteq Y \wedge A_{\exists} Y \subseteq \overline{Y}$

hint instantiate Y to $T_{\exists}^{*} X$

$\Rightarrow T_{\exists}^{*} T_{\exists}^{*} X \subseteq T_{\exists}^{*} X \wedge A_{\exists} T_{\exists}^{*} X \subseteq \overline{T_{\exists}^{*} X}$

hint $T_{\exists}^{*} T_{\exists}^{*} X \subseteq T_{\exists}^{*} X = \textbf{true}$

$\equiv A_{\exists} T_{\exists}^{*} X \subseteq \overline{T_{\exists}^{*} X}$

PROPOSITION 18 X is Internally Consistent is equivalent to there exists some Y subset of X such that Y is T-closed and Y is not in conflict with X.

$\Rightarrow$ **forward direction**

by $A_{\exists} T_{\exists}^{*} X \subseteq \overline{X}$

hint $T_{\exists}^{*} T_{\exists}^{*} X \subseteq T_{\exists}^{*} X = \textbf{true}$

$\equiv T_{\exists}^{*} T_{\exists}^{*} X \subseteq T_{\exists}^{*} X \wedge A_{\exists} T_{\exists}^{*} X \subseteq \overline{X}$

hint abstract over $T_{\exists}^{*} X$

$\Rightarrow \exists Y . T_{\exists}^{*} Y \subseteq Y \wedge A_{\exists} Y \subseteq \overline{X}$

$\Leftarrow$ **backward direction**

by $\exists Y . T_{\exists}^{*} Y \subseteq Y \wedge A_{\exists} Y \subseteq \overline{X}$

hint instantiate Y to $T_{\exists}^{*} X$

$\Rightarrow T_{\exists}^{*} T_{\exists}^{*} X \subseteq T_{\exists}^{*} X \wedge A_{\exists} T_{\exists}^{*} X \subseteq \overline{X}$

hint $T_{\exists}^{*} T_{\exists}^{*} X \subseteq T_{\exists}^{*} X = \textbf{true}$

$\equiv A_{\exists} T_{\exists}^{*} X \subseteq \overline{X}$

B. Trivial Examples Using MACE4

The simple examples used in chapter 5 to illustrate the translation of Dungian argumentation systems into Boolean Networks are used here to illustrate the use of MACE4 for model finding for such systems.

Example 5.4(i) has the translation

- $a \Rightarrow \neg b$,
- $b \Rightarrow \neg c$,
- $c \Rightarrow a$,
- $b \Rightarrow$ **false**,
- $a \Rightarrow$ **true**.

Using the MACE4 equivalents given in the following table:

operation	**symbolic**	**MACE4 input**
implies	$a \Rightarrow b$	`a -> b`
not	$\neg a$	`-a`
and	$a \wedge b$	`a & b`
or	$a \vee b$	`a	b`
true	**true**	`$T`
false	**false**	`$F`

gives the MACE4 input:

```
% inhibition

a -> -b.
b -> -c.

% feedback

a -> $T.
b -> $F.
c -> a.
```

For which MACE4 generates the assignments:

```
% number = 1
 a : 0
 b : 0
 c : 0

% number = 2
 a : 1
 b : 0
 c : 0

% number = 3
 a : 1
 b : 0
 c : 1
```

Continuing with the example 5.5(i) in the in the model checker's language gives:

```
 % inhibition

a1 ->  - b.
a2 ->  -b.
b  ->    -c.

% feedback

a1 -> $T.
a2 -> $T.
b -> $F.
c ->  (a1 | a2).
```

Giving assignments:

```
 % number = 1
 a1 : 0
 a2 : 0
 b  : 0
 c  : 0

% Interpretation of size 2
 a1 : 0
 a2 : 1
 b  : 0
 c  : 0

% number = 3
 a1 : 0
 a2 : 1
 b  : 0
 c  : 1

% number = 4
 a1 : 1
 a2 : 0
 b  : 0
 c  : 0

% number = 5
 a1 : 1
 a2 : 0
 b  : 0
 c  : 1

% number = 6
 a1 : 1
 a2 : 1
 b  : 0
 c  : 0

% number = 7
 a1 : 1
 a2 : 1
 b  : 0
 c  : 1
```

C. An Outline of the Theory of Information Security via Galois Connections

Confidentiality is defined either by semi-operational models, such as Bell-Lapadula [7, 8], or by abstract systems properties, such as Goguen and Meseguer's Non-Interference property [43], and its subsequent developments (see, for example, Peter Ryan's tutorial presentation "Mathematical Models of Computer Security" [96]).

Typically, in a model based account, there is an assignment of labels to pieces of information coding the information's classification, and an assignment of labels to individuals representing the individual's clearance. Access is modelled by a state transition system with states representing which individuals have access to what information. The system is constrained to follow transition rules obeying some invariance relating the classification of information to the clearances of individuals with access to that information. The constraints capture an abstract notion of a confidentiality policy for handling the information. Confidentiality is maintained if all system transitions obey the constraints.

The abstract systems property approach steps back a further level from this by not explicitly modelling the transition system. The system is regarded as some relation connecting inputs, carrying information at various levels of classification, to outputs, also at various levels of classification. That is to say, the notion of classification is associated with the systems inputs and outputs. Individuals can be identified with the outputs to which they have access[1]. The system property approach identifies confidentiality with the lack of information flow between selected inputs and selected outputs[2]. Different system properties are defined by taking different definitions of information flow.

A simple generalised view of the system properties approach can be obtained via Galois Connections. We may think of a system as an n-ary relation R that connects n *information spaces*, each space representing an input or output of the system. That is, R is a relation over n information sources S_i, i.e. $R : \mathbb{P}(\times_{i=1}^{n} S_i)$. We

[1] This is because we are discussing confidentiality. If we were to consider integrity as well we would need to identify individuals with the set of inputs and outputs to which they have access.

[2] Actually this view of inputs and outputs is not strictly accurate. Consider a cryptographic clock with no input and two outputs *high* and *low*. Output *low* will be the time as measured by the clock and output *high* is a secure digest of the time and a password creating a unique sequence of one-time-passwords. Any individual may read the time but only highly cleared individuals may read *high*. The system is secure if there is no information flow to *low* about the *high* output. We should regard inputs and outputs as a naming heuristic rather than as a definition of role in the systems property approach. Of course, the real problem here is that of finding a suitable definition of information flow.

now select two, disjoint, index sets $I, J \subseteq \{1, \ldots, n\}$, with $I \cup J = \{1, \ldots, n\}$, and consider the possibility of information flow between the subspaces $A = \times_{i \in I} S_i$ and $B = \times_{j \in J} S_j$.

This all amounts to considering a binary relation R' between the two spaces A and B. For concreteness we will say R' is from A to B. To simplify matters a little we will assume that R' is total in each direction. A property on a (sub)space is defined as a subset of the (sub)space. The observation of a property on a subspace of a system is the determination that the state of the system falls within the region of the subspace defining that property. Information flow between spaces A and B occurs when the observation of a property on B (i.e. that the system is in some particular region of B) is sufficient grounds to conclude that some particular property also holds on subspace A (i.e. we can determine a property of the A projection of the system from information obtained about the B projection of the system).

Specifically a property $A' \subseteq A$ is observable by observation $B' \subseteq B$ if $\breve{R}'^{\forall} A' = B'$.

That is, if the system R' is observed to be in region B' of B then the only elements of A which correspond to this situation are in A'. Note we do not assume that it is the case for every element of A'. Rather A' contains an *interior* A'' equal to $\breve{R}'_{\exists} B'$, i.e. equal to $\breve{R}'_{\exists} \breve{R}'^{\forall} A'$.

A property of A' is unobservable if $\breve{R}'^{\forall} A' = \emptyset$[7] (in which case the interior of A' is also empty). In some cases an observer may also be interested in knowing that some property definitely does not hold. A property A' is negatively observable when $\overline{A'}$ is observable, and negatively unobservable when $\overline{A'}$ is unobservable. If A' is both unobservable and negatively unobservable we will say it is completely unobservable. If A' is completely unobservable then no observation in B can discriminate between A' and $\overline{A'}$ is A.

One interesting aspect of this characterisation is that it relates properties of spaces i.e. observability depends on the chosen property of the space. Complete non-interference is obtained by generalising over all properties, that is: $\forall A' \subset A . \breve{R}'^{\forall} A' = \emptyset$. In which case no observation on B can reveal anything about A. Dualising the notion of observability to controllability creates the corresponding theory of integrity i.e. ensuring that a particular property A' holds on A is a sufficient condition to know that a particular property B' holds on B. Integrity requires that A is unable to control B. Taken together, observability and controllability give an explicit notion of correlation between spaces in which knowing that a property A' holds on A is the same as knowing that some particular property B' holds on B *and vice versa*. If A is an *input space*, meaning that it is possible to externally select which point the system occupies in A, and B is an *output space* meaning that it is possible to observe the whole of B, then this notion of correlation through R may legitimately be called *communication channel*, in that a signaller with access to A may send a message to an observer of B by coding using means of pre-agreed properties.

[7] That is, the unobservability of a property corresponds to: if the system is in A' then it may simultaneously be in any part of B.

Boolean Coherence: Does it make sense?

W. T. Harwood, J. A. Clark, J. L. Jacob

University of York
Department of Computer Science

Abstract. We continually face the problem of making sense of the world by resolving conflicting reports from multiple sources of information. This is particularly so if we attempt to formulate Qualitative Safety[1] Arguments. Traditional logic offers little to assist in this process.

In every day reasoning we usually assume that, without information to the contrary, we should use all information from all sources and that "information to the contrary" is the presence of an inconsistency between the sources. In order to resolve these conflicts we must make use of additional information which gives preference to one source of information over another when conflict arises between sources.

The suggestion put forward in this note is that it is possible to reach a 'best' conclusion by taking the most coherent theory that respects the preference ordering on sources and that this theory is the maximal consistent theory with respect to the ordering. This process is parameterized by an underlying notion of logic that provides the notions of consistency and entailment. This notion is straightforward when applied to the standard notion of a strict preference ordering but is a little more involved when we consider partially ordered preferences.

1 Introduction

A safety case documents the argument that links evidential claims to a safety claim that they support. The evidential claims are propositions whose truth is supported by evidence. The safety claim is a direct or indirect logical consequence of the evidential claims. This is essentially the view set out by e.g. Bishop and Bloomfield [3] or Wilson, Kelly and McDermid [14] (although both include additions and extensions to this basic view). If this was all there was to consider, the world would be simple and standard logical reasoning would allow one to take the evidential claims and prove (or not) the safety claim. Indeed, Rushby considers how a large system of claims might be formalized in a way that could be handled by existing theorem proving tools [11]. The problem, however, is that the real world is not so simple. Evidential claims need to be selected from a collection of potentially conflicting claims that arise from different sources of information and that are supported by varying degrees of

[1] Here, and throughout, safety refers to the real world safety of a system, as opposed to the technical notions of safety (and liveness) introduced into program verification by Lamport [8].

evidence. Moreover the notion of degree, or strength, of evidence is far from straightforward. In general we are neither neutral to what is asserted as an evidential claim nor neutral about the source from which the claim arises. We rate the plausibility of a claim by such assessments as the degree to which it accords with our own experience, the degree of bias we might think is being expressed, the reproducibility of results, the experience of an individual making a claim, the methodology (or lack of it) that lead to the claim, etc.

A safety case will simply be that some *suitably selected* set of the evidential claims entails the safety claim. The issue is what does *suitably selected* actually mean?

The approach taken here is to at least partially formalize these notions in a framework we call Boolean Coherence[2]. It is related to Rescher and Manor's paraconsistent logic [10] and Default Logic [9] in the way that it deals with the presence of inconsistencies, and is strongly related to Prioritized Default Logic (see, for example, [2, 5, 6]) in its use of a preference relation to decide which assertions should be considered active during reasoning. Unlike these logics however, our concern is with selecting a consistent subset of the evidentially supported claims rather than in defining a different notion of entailment. The approach is parameterized by an underlying notion of logic which can be varied, but for this paper we will assume the underlying logic to be classical propositional logic. Entailment comes from the underlying logic and is used to show that the claims either hold, or fail to hold, with respect to this selected subset. This has the effect of making our notion of entailment more conservative than those used either in systems of paraconsistent or default logic.

To select one claim over another requires the use of additional information about the claims. We require each claim to be labeled by a *source* and that a *preference ordering* is imposed over sources.

A *source* is a formal notion reflecting whatever it is that supports the veracity of the claim (e.g. an individual making the claim, a process that produces a result, etc.), and the *preference ordering* over sources expresses the idea that some sources reflect a higher strength of evidence than others. The selection rule is that we start with the strongest sources and progressively include information from weaker sources. When contradictions arise between sources ordered by the preference relation then we preserve stronger sources in favour of weaker sources. When contradictions arise between claims not strictly ordered by the preference relation we remove the *weakest set* of sources that removes the contradiction, treating both sides of the contradiction symmetrically.

Consider a simple hypothetical example: A company intends to place a radio transmitter mast in a location near a school. Legislation permits the company to operate in one of two bands, Band X and Band Y. Recent medical opinion from extensive experimentation on mice is that Band X is unsafe. An expert biophysicist believes that if Band X is unsafe then the same is true of Band Y, although he has no experimental evidence to back this up. A technical expert is willing to testify that the company operates in Band Y. We might

[2] In contradistinction to Bayesian Coherence as discussed in, e.g. [4]

consider the ordering of plausibility of this evidence as: legislation is the most definite fact, the medical expertise and the technical expertise are on a par and that the biophysicist is expressing an opinion which is less well founded i.e. legislation > medicalExpertise = technicalExpertise > biophysicist. The conclusion we arrive at is that it is unsafe to place the transmitter mast given the best evidence available. If now, however, additional evidence from a medical expert is received to the effect that the same experiments that were performed on mice using Band X were performed on mice using Band Y and there were no ill effects, then we would revise our conclusion even if the biophysicist still holds the same opinion about the relation between Band X and Band Y. The reason we revise our assessment is that the new medical claim is stronger than the biophysicist opinion because it is backed by experimental evidence. In essence the overall safety case can be expressed as the hypothesis 'safe' follows from the most plausible theory we can construct given the available sources of evidence and our relative evaluations of the plausibility, reliability or trustworthiness of the evidence. Using a fairly self evident formalism the overall safety case can expressed as:

```
legislation >
medicalExpertise =  technicalExpertise >
biophysicist  |

legislation: (BandX & ~BandY) + (~BandX & BandY),
medicalExpertise: BandX => ~safe,
medicalExpertise: BandY => safe,
biophysicist: (BandX => ~safe) => (BandY => ~safe),
technicalExpertise: ~BandX

hypothesis
safe
```

The ordering gives the reason that we reject the evidence of the biophysicist, i.e. that, in the context of the other claims, it contradicts 'stronger' claims.

Our goal is to formalize such reasoning to enable the capture of all of the claims, whether used or rejected, in constructing the safety case.

2 Formalisation

The reasoning process can be viewed as finding a maximal consistent set of sources that is also a maximal set in an ordering obtained when the source ordering is extended to an ordering on sets of sources. This extension to sets of sources should obey two simple conditions:

- Given two sets of sources, A and B, any sources they have in common cannot help decide between them.

- Given two sets of sources, A and B, with no sources in common, A is greater than B if for every element b of B, A has some element greater than b

That is, when we ignore the elements that A and B have in common, A is bigger than B if, whichever element of B we look at, we can always find a bigger element of A.

We formalize this as a relation over non-empty sets, by: Let $\succ$ be the ordering on sources, which is a strict partial order (i.e. irreflexive and transitive), possibly obtained as the strict part of a non-strict partial order (i.e. reflexive, transitive and anti-symmetric), then, for non-empty, distinct, sets A and B[3],

$$A \sqsupset B \equiv \forall b \in B \setminus A.\ \exists a \in A \setminus B.\ a \succ b$$

and otherwise false.

If $\succ$ is a strict total order then there is only one maximally consistent set that is also maximal in this ordering $\sqsupset$. If $\succ$ is a strict partial order then this is no longer the case and there may be none, one or many maximally consistent sets that are 'largest'. That is, we have a collection of sets that are incomparable under the ordering $\sqsupset$ and are inconsistent with one another. In this case we opt to take the common elements of all the maximally consistent sets that are also maximally in $\sqsupset$.

Let P be a propositional language and let L be a set of labels denoting sources. The set of pairs $L \times P$ is the labeled propositional language generated by L and P and we write elements of $L \times P$ as $l : p$ where $l \in L$ and $p \in P$. We define the projections on sets of pairs $prop(S) = \{p \mid \exists l.\ l : p \in S\}$, $lab(S) = \{l \mid \exists p.\ l : p \in S\}$ and the selective projection, for a set of labeled propositions S and a set of labels l, $S \circ l = \{p \mid x : p \in S \wedge x \in l\}$.

We assume P is equipped with a consistency predicate $CONS\ Q$ which determines for each $Q \subseteq P$ whether or not Q is consistent and an entailment relation $\vdash$ which determines if a set $Q \subseteq P$ entails a given p, element of P[4]. Given a set of labeled propositions, S, we define $cons\ S$ as the set of all consistent subsets of S by[5]:

[3] We should note that because this relation is only used over maximally consistent sets to decide on which of two sets makes a 'better' choice with respect to resolving inconsistencies, there is a degree of leeway in the exact relation that could be used. We may also note that the relation we have used is closely allied to the Hoare ordering used in defining the lower, or Hoare, power domain.

[4] If we restrict ourselves to classical logic we may avoid having both $CONS$ and $\vdash$ as they are inter-definable. However, they are not necessarily so in a more general setting. Moreover, one may be tempted to think of the setup with both relations as a Scott information system [13] but this is not necessarily the case, as the underlying logic not obey the axioms of information systems. For example, the underlying logic could fail transitivity of entailment for consistent sets.

[5] We treat the collection of assertions with the same label as if they were a single conjunctive assertion with that label.

$$cons\ S = \{S \circ l \mid l \subseteq lab(S) \wedge CONS(S \circ l)\}$$

Next we define the maximal sets under subset ordering ($\supset$) and the extended preference ordering ($\sqsupset$):

$$max_{\supset} X = \{M \in X \mid \neg \exists N \in X . N \supset M\}$$

$$max_{\sqsupset} X = \{M \in X \mid \neg \exists N \in X . lab(N) \sqsupset lab(M)\}$$

and these are used to define the set of alternative theories each of which is maximal in the extended preference ordering:

$$alternatives\ S = max_{\sqsupset}(max_{\supset}(cons\ S))$$

Finally the most plausible theory is defined as the common elements of the alternative maximal theories:

$$plausible\ S = \begin{cases} \emptyset & \textbf{if } alternatives\ S = \emptyset \\ \bigcap(alternatives\ S) & \textbf{otherwise} \end{cases}$$

The effect of the first maximization, by $max_{\supset}$, is to take maximally consistent sets. To aid the understanding the second maximization, $max_{\sqsupset}$, we introduce the notion of a *conflict* within a set of propositions. A conflict in a set S is a minimal contradiction within S, i.e. a conflict is a set $C \subseteq S$ such that $\neg CONS(C)$ and $\forall C' \subset C.\ CONS(C')$. Let $conflicts(S)$ be the set of all conflicts in S. We will say $\succ$ uniquely resolves the conflicts of a set S if every $C \in conflicts(S)$ has a minimum element under this ordering.

Given a maximally consistent set A and a conflict C in A, then at least one $c \in C$, is not in A. We say A excludes c. For maximally consistent sets, A and B, $A \sqsupset B$ means that A excludes less preferred elements of conflicts than does B. If $\succ$ uniquely resolves the conflicts of a set S then there is a single largest (under $\sqsupset$) maximally consistent subset of S. If $\succ$ does not uniquely resolve the some conflicts of S then there are multiple maximally consistent sets under $\sqsupset$ that differ in the elements they exclude of each conflict. As there is no preference between these resolutions we cannot decide between them and so reject all the alternative resolutions in favour of the common part of the maximally consistent sets.

Returning to the example above and applying this definition of plausible using a simple Boolean Coherence calculator program, the plausible theories for “before” and “after” the additional medical information is obtained are:

```
medicalExpertise: BandX => ~safe
biophysicist: BandX => ~safe => BandY => ~safe
technicalExpertise: ~BandX
legislation: BandX & ~BandY + ~BandX & BandY
```

and

```
legislation: BandX & ~BandY + ~BandX & BandY
medicalExpertise: (BandX => ~safe)   &
                  (BandY => safe)
technicalExpertise: ~BandX
```

A hypothesis holds if it is entailed by the most plausible theory. In our example case this gives either `~safe` or `safe` respectively.

3 A Comparison with Some Other Logics

We briefly compare this approach to the paraconsistent logic of Rescher and Manor and prioritized default logic.

Rescher and Manor define two closure operators derivable from the set of maximally consistent sets in the presence of inconsistency, here modified to deal with labeled propositions:

$$\begin{aligned} \mathit{strong}(S) &= \bigcap(\mathit{Th}(\mathit{max}_{\supset}(\mathit{cons}\ S))) \\ \mathit{weak}(S) &= \bigcup(\mathit{Th}(\mathit{max}_{\supset}(\mathit{cons}\ S))) \end{aligned}$$

with $\mathit{strong}(S) \subseteq \mathit{weak}(S)$.

A proposition, p, is true in $\mathit{strong}(S)$ if it follows from every maximally consistent subset, i.e. it may be true for different reasons in each maximally consistent set. A proposition, p, holds in $\mathit{weak}(S)$ if it follows from any maximally consistent set. In this case p may be true in some maximally consistent sets and $\neg p$ may be true in others. This does not necessarily mean that all propositions are true in $\mathit{weak}(S)$ since these contradictions do not collide inside a *Th*-closure.

If we restrict our attention to situations where $\succ$ uniquely resolves all conflicts in S then $\mathit{strong}(S) \subseteq \mathit{Th}(\mathit{plausible}(S)) \subseteq \mathit{weak}(S)$.

Prioritized default logic is an extension of Reiter's default logic. We give a brief sketch of Reiter's logic and its extension to prioritized default logic.

Reiter's default logic is constructed by extending classical logic with *default rules* of the form $(\frac{\alpha:\beta_1,\ldots,\beta_n}{\gamma})$, where α is called the prerequisite, β_i the justifications and γ the conclusion, of the default rule. A rule is *active* if α is true in the current theory and each β_i is consistent with the current theory. If a rule is active then the current theory can be extended by γ. If T is a classical theory (deductively closed collection of propositions under classical inference) and D is a collection of default rules, a theory extension of the default presentation (D,T) is generated by repeatedly, whilst possible, non-deterministically selecting an active default rule and applying it to obtain a new theory T' in which all consequences of the conclusion of the rule are added to the theory. When it is no longer possible to continue because there are no more active rules to apply, if every justification of every default rule used in generating T' is consistent with T', then T' is a default extension of T. A conclusion p follows *skeptically* from (D, T) if p is true in all default extension of (D, T) and p follows *credulously* from (D, T) if p is true in some default extension of (D, T). An extension to these approaches is to say a proposition p is true *preferentially* if it holds in all *best preferred extension* defined by a selection rule[7].

Restricting default rules to the form $(\frac{:p}{p})$ gives a logic essentially the same as Rescher and Manor's with skeptical consequence corresponding to strong closure and credulous consequence corresponding to weak closure.

Prioritized default logic extends default logic by allowing the specification of priorities between default rules. A priority relation is a strict partial order. Priorities may be used in at least two distinct ways: default extensions are built using the highest priority active default rules available at each step of the non-deterministic iteration; or default extensions are built in the standard way except when a conflict arises between rules, in which case the highest priority rule is applied. In either case, if there are multiple extensions, default entailment may be taken skeptically, credulously or preferentially.

Our logic closely corresponds to a prioritized default logic with rules restricted to the form $(\frac{:p}{p})$ and partially ordered with the ordering used to resolve conflicts and the semantics taken as skeptical entailment. It differs in the entailment when the order relation does not fully resolve all conflicts in a set of propositions because we take a more conservative entailment. A proposition p is only entailed if it can be entailed from the same premises in all default extensions i.e. it follows from the common part of the default extensions.

4 Conclusion

The logic presented here captures a notion of entailment from the most plausible theory given some notion of ordering of the sources of information. Considering the most plausible theory drawn from all the available evidential claims causes us to be explicit in rejecting information from sources. Information is only rejected if there is more plausible, that is more preferred, information. Rejection of information therefore becomes a matter of providing explicitly the contrary case together with the assertion that the contrary case is more plausible than the rejected information thereby making the overall safety case more explicit.

The example discussed here uses classical propositional logic but it should be clear from the formalization that any logic that offers a notion of consistency and entailment can be used in its place. Practical calculation proceeds by using the ordering to decide how to attempt to extend consistent sets of information and backtracking when inconsistencies are encountered. The ability to perform the calculation is limited by the complexity of the satisfaction problem, which in the current case is NP. This problem notwithstanding, it is interesting to consider the use of logics, such as a conditional logic or relevance logic, to better reflect the information relation between assertions. One may also consider how much *effort* is put into finding the most plausible theory. We may, for example, consider limiting consistency checking by some computational limit, reflecting the idea that there is only bounded foresight when considering how claims interact. Such changes have the potential for formally capturing more of the practical reasoning of real world a safety cases.

5 Reviewer Acknowledgments

We thank the anonymous reviewers for their insights, comments and questions. We have attempted to address most of them without substantially changing the flow of the paper. However, one observation has gone unaddressed: there is a connection between the approach taken here and the field of multi-valued model checking (c.f. [12]). This is an interesting observation as there is a significant connection between paraconsistent and relevance logics, and multi-valued logics, particularly the 4-valued logic of Belnap [1]. The approach in [12] seems to extends Belnap's framework to multiple participants suggesting an approach to multiple sources we had not considered. We hope to investigate this suggestion in a later paper.

6 Sponsorship Acknowledgement

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

References

1. Alan Ross Anderson and Nuel Belnap. *Entailment: v. 2: Logic of Relevance and Necessity*. Princeton University Press, 1992.
2. Franz Baader and Bernhard Hollunder. Priorities on defaults with prerequisites, and their application in treating specificity in terminological default logic. *J. Autom. Reasoning*, 15(1):41–68, 1995.
3. Peter Bishop and Robin Bloomfield. A methodology for safety case development. In *Safety-Critical Systems Symposium*. Springer-Verlag, 1998.
4. Luc Bovens and Stephan Hartmann. *Bayesian Epistemology*. Oxford University Press, 2003.
5. John Horty. Defaults with priorities. *Journal of Philosophical Logic*, 36(4):367–413, 2007.
6. Yarden Katz and Jennifer Golbeck. Social network-based trust in prioritized default logic. In *AAAI*. AAAI Press, 2006.

7. Sarit Kraus, Daniel J. Lehmann, and Menachem Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *CoRR*, cs.AI/0202021, 2002.
8. Leslie Lamport. Proving the correctness of multiprocess programs. *IEEE Trans. Software Eng.*, 3(2):125–143, 1977.
9. Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.
10. N. Rescher and R. Manor. On inference from inconsistent premises. *Theory and Decision*, pages 179–217, 1970.
11. John Rushby. Formalism in safety cases. In *Making Systems Safer*, pages 3–17. Springer, 2010.
12. Mehrdad Sabetzadeh and Steve Easterbrook. Analysis of inconsistency in graph-based viewpoints: A category-theoretic approach. *Automated Software Engineering, International Conference on*, 0:12, 2003.
13. Dana S. Scott. Domains for denotational semantics. In Mogens Nielsen and Erik Meineche Schmidt, editors, *ICALP*, volume 140 of *Lecture Notes in Computer Science*, pages 577–613. Springer, 1982.
14. S P Wilson, T P Kelly, and J A McDermid. Safety case development: Current practice, future prospects. In *1st ENCRESS/12th CSR Workshop*. Springer, 1995.

Boolean Coherence and the ACH method

W. T. Harwood J. A. Clark, J. L. Jacob
University of York, Department of Computer Science
R. I. Young, DSTL, UK

Abstract—**Richards Heuer's book, "The Psychology of Intelligence Analysis", sets out an approach to Intelligence Analysis based on comparing Alternative Competing Hypothesis (ACH). Heuer's work is expressed in terms of an informal logic of performing such comparisons. Previous authors have formalized Heuer's approach in a probabilistic setting. Here we set out an alternative formalization in a non-probablitistic setting which only requires the information sources to be partially ordered with respect to their relative trustworthiness or reliability. This approach avoids the need to give numeric estimates of trustworthiness or reliability and, as such, may be better suited to many real life situations in which it is not possible to obtain meaningful estimates these quantities. The analysis approach has been automated and we give a short example of the resulting analysis.**

I. Introduction

We consider the problem of formalizing the reasoning process that supports Richard Heuer's "The Psychology of Intelligence Analysis" [11].

Heuer presents a picture of Intelligence Analysis as an essentially Popperian process of creating a complete set of alternative hypotheses about a situation and then attempting to systematically refute them using the available evidence ([13, 14]). Hypotheses that survive the refutation process are then regarded as possible explanations for the situation. The two main difficulties in the process are determining what constitutes a complete set of hypotheses, and how to deal with the fact that the available evidence to be used in the refutation process is often a conflicting set of reports about the state of the world.

Our approach to the first problem is to regard a hypothesis as a binary condition and the refutation process as one of determining the consistency of the both the hypothesis and its negation with the available data. This means that a hypothesis may be in one of three states: both it and its contradiction may be consistent with the data; it may be consistent with the data and its contradiction inconsistent with the data; or it may be inconsistent with the data but its contradiction is consistent with the data. An Intelligence Analyst forms a set of hypotheses by forming a boolean vector of the characteristics of interest e.g. she might characterize a potential attacker by a list of binary attributes and the reasoning process will assign one of the three states to each attribute.

The second difficulty of dealing with potentially inconsistent information is addressed by putting the information sources into a hierarchy representing the relative trustworthiness or reliability of the sources. Generally we refer to the relation of relative trustworthiness or reliability of the sources as the *preference* relation over the sources. Our reasoning process attempts to build the most plausible theory possible given the available information and the preference relation over the sources. Once a most plausible theory has been generated each hypothesis, and its negation, are checked for consistency with the theory. The rule for using information in building the theory is to use all information unless a contradiction arises between sources. When a contradiction arises, the more preferred source is taken over the less preferred source and all information from the less preferred source is rejected. Although this may seem extreme it is easy to express sources where a rejection of one piece of information does not cause all the information from the source to be rejected by splitting a single source into multiple logical sources.

The mechanism of using sources with preference relations offers a quite flexible model of ranking information. For example, we may combine an intrinsic notion of trustability of a physical source with a measure of the strength of evidence associated with a statement obtained from a source. For concreteness sake let us suppose the physical sources are P_1 and P_2 with the intrinsic trustworthiness as $P_1 \succ_t P_2$ and the strength of evidence being *strong* and *weak* with *strong* $\succ_e$ *weak*. We may then create logical sources $(strong, P_1)$, $(weak, P_1)$, $(strong, P_2)$ and $(weak, P_2)$ and order them by the lexicographic ordering $(\succ_e, \succ_t)$ i.e.

$$(x, p) \succ (y, q) \equiv x \succ_e y \vee (x = y \wedge p \succ_t q)$$

That is, when there is equal strength of evidence we rely on the intrinsic trustworthiness of the source but when one set of evidence is stronger than another, the strength of evidence determines which piece of information we prefer.

Finally we note Heuer's discussion is framed as essentially static in that it does not specifically address the dynamics of belief revision as the available information changes. We do not regard this as a fundamental issue in Heuer's work, rather it is seen as a presentational issue i.e. one computes the best set of hypothesis given the available information. But new information may cause non-monotonic revision of the hypotheses evaluation. This interpretation leads intelligence evaluation to be seen as a dynamic process in which there is no 'right answer' but rather only a best answer given the currently available information and set of assumptions.

We continue the paper with a small, fictitious, example of how an intelligence analysis might evolve, followed by a brief overview of the theory behind the reasoning process. The theory itself has been implemented in a small calculator program to permit experimentation with examples.

II. An Extended Example

We consider a simple situation in which an analyst has various sources of information about a possible terrorist attack. Firstly, she has specific knowledge of attacks acquired over the years which places bounds on her expectations of the terrorist activities. This knowledge is incomplete and may be incorrect but it is the most certain knowledge she has. Then there is information that she may obtain from sources such as intercepts, police reports, sightings at transit points and informers. This information has various degrees of credibility and can be arranged in a hierarchy. The general rule is that she will consider all information as true but when a conflict arises between two sources she will discard the less credible source in favour of the more credible. Note that this will mean a single conflict causes all the information provided by the least credible source to be disregarded. If this is an issue for a particular source, e.g. the source is regarded as more credible for certain types of information than for others, this is handled by splitting the source into multiple independent sources, each providing a part of the sources information.

Our analyst has a set of *background* assumptions built up from experience:

- That potential terrorists fall into two categories, Professional and Amateur. Professionals are further divided into Career terrorists and Disposable terrorists. Career terrorists carry out repeated acts of terrorism whereas Disposable terrorists carry out suicide missions. Professionals tend to operate with support teams and the identification of the presence of a support team is sufficient to indicate an attack will be professional.
- The modes of attack that are available are Sniper, Bomb and Mortar, with Bomb divided into Placeable, Car Bomb and Suicide Bomb. A career terrorist may attack as a Sniper, with a Placeable Bomb or with a Mortar.
- A Disposable terrorist will attack with a Suicide Bomb or a Car Bomb (which is often regarded as a form of Suicide Bomb). And an Amateur will act as a Sniper or use a Suicide Bomb.
- The only true distance attack option is the Mortar and this is only used by Career terrorist
- Amateurs will use homemade explosives, whereas professionals will steal or purchase explosives, or purchase explosive precursors (not necessarily legally).
- The quantity of materials involved will indicate the likely size of an explosive device. So reports of theft or purchase of small quantities of explosive will indicate Placeable or Suicide Bombs, whereas large quantities will indicate Car Bombs. Purchase of large quantities of precursors will indicate a Car Bomb but purchase of small quantities of precursors is likely to go unnoticed and also is unlikely to be indicative (since by their very nature they can be purchased for legitimate ends).
- Career terrorists always have an escape plan, Amateur and Disposable do not need one.

For this particular example we will assume that the analyst believes there will only be a single attack (as opposed to the possibility of multiple simultaneous attacks).

In addition to these background assumptions, the analyst also has some weak *default assumptions* that, given no evidence to the contrary, the analyst takes as good working hypothesis. In this case:

- The attack will be carried out by an amateur.

The analyst partially orders the information sources by reliability and trustworthiness. In doing so she must also assess where her own assumptions sit with respect to other sources of information. For example, she may regard intercepts and police reports as definite pieces of information that are more reliable than her own assumptions, whereas she may regard sightings as less reliable, and informers as simply less trustworthy than her own background assumptions (but as a more reliable guide than her default assumptions). In examining her own background assumptions she might find that they fit into logical groupings such that if any item in the grouping was to be contradicted by a preferred source then she would give up the entire group of assumptions. In this example we divide the analyst's background assumptions up into assumptions about the weapons that might be used in an attack (WEAPONS), assumptions about the type of attacker (ATTACKER_TYPE), assumptions about the preferred weapons associated with a type of attacker (ATTACKER_PREFERENCES) and assumptions about whether there is a single of multiple type of attack (SINGLE_ATTACK) and, if the attack uses explosives, whether or not there is a single type of bomb involved (SINGLE_BOMB_TYPE). The preference ordering is illustrated in figure 1 (there may also be a set of definitions which is taken as trustworthy, one might say, by definition).

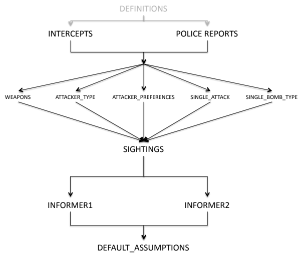

Fig. 1. Example Ordering

The analyst also forms a collection of hypothesis of interest, in this case:

- Type of attacker:
  - Amateur,
  - Professional:
    - Career,
    - Disposable
- Type of attack:
  - Bomb:
    - Placeable Bomb,
    - Suicide Bomb,
    - Car Bomb
  - Sniper,
  - Mortar

Our example unfolds in seven time steps as reports come in from various information sources

1) An initial intercept reports a bomb threat. At this stage the *default assumption* that the attacker will be an amateur restricts the alternative attacks to suicide bomb or car bomb.
2) A sighting of support personnel causes the deduction that the attacker is a professional. This overrides the default assumption and replaces the attacks by placeable bomb or suicide bomb. At this stage the attacker may be a career terrorist or a disposable terrorist.
3) A police report of a small theft of explosives comes in. This is confirmatory evidence that it is a career terrorist and does not alter the set of the most plausible hypotheses.
4) Informer 1 reports that the attack will be carried out by a sniper. This is contradicted by a more preferred information source, intercepts, and so is discarded.
5) Informer 2 reports that there is no escape plan. This means that the attacker is a disposable professional and the attack is a suicide bomb.
6) Intercepts reports the discovery of an escape plan. This causes the reports from informer 2 to be discarded because the intercepts source is preferred over informer 2. The attacker is re-categorized as a career professional and the attack is re-categorized as a placeable bomb.
7) Intercepts reports that a simultaneous bomb threat and sniper attack will take place. This undermines the

analyst's assumption of a single attack and backs up informer 2's report. This causes sniper to be added to the attacks and opens the possibility of a mortar attack as well (since the analyst's exclusion principle has been discarded).

Table I sets out this timeline of reports from the various sources and the evolution of the hypothesis set. Appendix A gives the formalization of the background assumptions, default assumptions and ordering used in the example.

TABLE I
Evolution of Hypotheses

TIME	1	2	3	4
SOURCE	INTERCEPTS	SIGHTINGS	POLICE REPORTS	INFORMER 1
REPORT	Bomb Threat	Support	Small Theft Explosive	Sniper
Amateur	+	-	-	-
Professional	-	+	+	+
Career	-	+/-	+/-	+/-
Disposable	-	+/-	+/-	+/-
Bomb Threat	+	+	+	+
Placeable Bomb	-	+/-	+/-	+/-
Suicide Bomb	+/-	+/-	+/-	+/-
Car Bomb	+/-	+/-	-	-
Sniper	-	-	-	-
Mortar	-	-	-	-
TIME	5	6	7	
SOURCE	INFORMER 2	INTERCEPTS	INTERCEPTS	
REPORT	No Escape Plan	Escape Plan	also Sniper	
Amateur	-	-	-	
Professional	+	+	+	
Career	-	+	+	
Disposable	+	-	-	
Bomb Threat	+	+	+	
Placeable Bomb	-	+	+	
Suicide Bomb	+	-	-	
Car Bomb	-	-	-	
Sniper	-	-	+	
Mortar	-	-	+/-	

Although the example presented here shows an evolution as information is obtained, it may still give the impression of being somewhat static in that the background theory, sources and order relation are given up front. This is not essential to the approach. There is no reason why the set of sources, background assumptions, default assumptions and the ordering should not be modified at each step in the analysis.

III. The Formal Model

Although throughout this paper we use classical propositional logic the formalization of the reasoning process is relatively independent of the underlying logic used. The requirement is that we have some suitable versions of the notion of consistency, so we can discuss what-conflicts-with-what, and of entailment, so we can discuss what-follows-from-what. We define the formal version of our reasoning process in terms of a language equipped with both of these notions. In classical logic these notions are inter-definable (e.g. $\Gamma \models \phi$ iff $\Gamma \cup \{\neg\phi\}$ is inconsistent) but we may consider other logics where the connection is weaker. Consider, for example, extending the idea of the inconsistency of a set of propositions to include an additional constraint, such as physical realizability. Such a condition may be used to rule out unacceptable situations e.g. those that imply traveling backwards in time.

Let P be a propositional language and let L be a set of labels denoting sources. The set of pairs $L \times P$ is the labeled propositional language generated by L and P and we write elements of $L \times P$ as $l : p$ where $l \in L$ and $p \in P$. We define the projections on sets of pairs, S, by $prop(S) = \{p \mid \exists l.\ l : p \in S\}$ and $lab(S) = \{l \mid \exists p.\ l : p \in S\}$ and the selective projection, for a set of labeled propositions S and a set of labels l, $S \circ l = \{p \mid x : p \in S \wedge x \in l\}$.

L is equipped with an strict partial order (i.e. transitive and irreflexive relation) $\succ$. P is equipped with a consistency predicate *consistent* Q which determines for each $Q \subseteq P$ whether or not Q is consistent and an entailment relation $Q \vdash p$ which determines if a set $Q \subseteq P$ entails a given p, element of P.

Our intuition on the order relation is that its fundamental role is to decide between sources when contradictions arise between them. That is, it gives us a criteria for the unacceptability of a source.

Given a set of labelled propositions S, let $I(S)$ be the set of minimal inconsistent subsets of labels of S, defined by:

$$I(S) = \{lab(i) \mid i \subseteq S \wedge \neg consistent(S \circ lab(i)) \wedge \\ \forall j \subset i. consistent(S \circ lab(j))\}$$

Given a minimal inconsistent set of labels, i, we say that the *causes* of the inconsistency are the minimal elements (under the ordering $\succ$) of the set. Minimal elements are defined as the the least elements of each of the maximal chains in an inconsistent set of labels i.

$$chains(i) = \{j \subseteq i \mid \forall x, y \in j.\ x \neq y \implies x \succ y \vee y \succ x\}$$

$$mchains(i) = \{j \in chains(i) \mid \forall k \in chains(i).\ j \not\subset k\}$$

$$causes(i) = \{x \mid \exists c \in mchains(i).\ x \in c \wedge \forall y \in c. x \neq y \implies y \succ x\}$$

Under this definition of the causes of a contradiction, if there is a least element in an inconsistent set of labels i, then this element is the cause of the contradiction. If, on the other hand, there is no single least element because not all elements are comparable, then the least elements of the incomparable chains are taken as the causes of the contradiction. In the case of all elements being mutually incomparable, all elements of the contradiction are taken as mutual causes (since we have no way of preferring one cause over another).

The complete set of all causes of inconsistencies in S is defined as the set of unacceptable labels in S :

$$unacceptable(S) = \bigcup\{causes(i) \mid i \in I(S)\}$$

We now define the acceptable elements of S as those sets that do not contain a cause of an inconsistency.

$$acceptable(S) = lab(S) \setminus unacceptable(S)$$

The most plausible theory of S is taken as the subset of S labeled by acceptable labels i.e. the largest set of elements of S that do not contain the causes of contradictions.

We briefly compare this semantics to the paraconsistent semantics introduced by Rescher and Manor [16]. Let $Th(S)$ be the deductive closure of a set of propositions. Then there are 4 closure operators derivable from the set of maximally consistent sets in the presence of inconsistency. Let $mcons(S)$ be the set of maximally consistent subsets of S defined by:

$$mcons\ S = \{S \circ l \mid l \subseteq lab(S) \wedge consistent(S \circ l) \wedge \\ \forall l'.\ l \subset l' \subseteq lab(S) \implies \neg consistent(S \circ l')\}$$

$$\begin{aligned}
\mathit{conservative}(S) &= \mathit{Th}(\bigcap(\mathit{mcons}(S)))\\
\mathit{strong}(S) &= \bigcap(\mathit{Th}(\mathit{mcons}(S)))\\
\mathit{weak}(S) &= \bigcup(\mathit{Th}(\mathit{mcons}(S)))\\
\mathit{inconsistent}(S) &= \mathit{Th}(\bigcup(\mathit{mcons}(S)))
\end{aligned}$$

Where *strong* and *weak* are Rescher and Manor's notions of strong and weak closures modified to deal with labeled propositions[1] and *inconsistent* is the obvious inconsistent set of all propositions (equally specified, less symmetrically, as $\mathit{Th}(S)$).

Clearly the closures are ordered by subset inclusion $\mathit{conservative}(S) \subseteq \mathit{strong}(S) \subseteq \mathit{weak}(S) \subseteq \mathit{inconsistent}(S)$. A proposition, p is in $\mathit{conservative}(S)$, if it follows from the same propositions in every maximally consistent set, whereas it is true in $\mathit{strong}(S)$ if it follows from every maximally consistent subset, i.e. it may be true for different reasons in each maximally consistent set. A proposition, p, holds in $\mathit{weak}(S)$ if it follows from any maximally consistent set. In this case p may be true in some maximally consistent sets and $\neg p$ may be true in others. This does not mean that all propositions are true in $\mathit{weak}(S)$ since these contradictions never collide inside a Th-closure. Finally all propositions hold in $\mathit{inconsistent}(S)$.

Adding labels and preferences changes the notion of a maximally consistent set to that of a maximally consistent set that does not contain unacceptable elements. Let us call the new notion maximally preferentially consistent, *mpcons*, defined as:

$$\begin{aligned}
\mathit{mpcons}\ S = \{S \circ l \mid l \subseteq \mathit{lab}(S) \wedge \mathit{consistent}(S \circ l) \wedge{}\\
\forall l'.\ l \subset l' \subseteq \mathit{lab}(S) \implies \neg\mathit{consistent}(S \circ l') \wedge{}\\
l \cap \mathit{unacceptable}(S) = \emptyset\}
\end{aligned}$$

Replacing *mcons* by *mpcons* in the definitions of the closure operators we obtain preferential versions of each operator, $\mathit{conservative}_p$, strong_p, weak_p and $\mathit{inconsistent}_p$, with the inclusion between operators maintained. If, additionally, we restrict our attention to situations where the set of labeled propositions, S, and the order relation over sources is such that $\mathit{causes}(i)$ is a singleton for each $i \in I(S)$[2] then $\mathit{conservative}(S) \subseteq \mathit{conservative}_p(S)$, $\mathit{strong}(S) \subseteq \mathit{strong}_p(S)$, $\mathit{weak}_p(S) \subseteq \mathit{weak}(S)$ and $\mathit{inconsistent}_p(S) \subseteq \mathit{inconsistent}(S)$. However, $\bigcup\ \mathit{mpcons}(S) = S \setminus \mathit{unacceptable}(S) = \bigcap \mathit{mpcons}(S)$, so the operators collapse into a single operator because the notion of unacceptability resolves the choices of resolving inconsistencies. Let us call this operator $\mathit{acceptable}_p$. Then we have, $\mathit{strong}(S) \subseteq \mathit{acceptable}_p(S) \subseteq \mathit{weak}(S)$.

IV. Conclusion

We have given a brief analysis of the ACH process and a logic that supports the ACH process in Intelligence Analysis. Currently we use a trivial program to perform the analysis based on a relatively inefficient tableaux reasoner and work with a small number of propositions. But the logic is well suited to automation by SAT solvers for finding contradictory sets with large numbers of boolean variables. We hope that future work will allow us to explore the utility of this form of reasoning in practical applications with large data sets. On the theoretical side there is much to be said, and even more to be explored, about the logic and its relation to other paraconsistent logics[3]. and preferential default logics[4]. This, we hope, will be the subject of another paper.

V. Acknowledgements

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

References

[1] Seiki Akama. Nelson's paraconsistent logics. *Logic and Logical Philosophy*, 7:101–115, 1999.

[2] F. G. Asenjo. A calculus of antinomies. *Notre Dame Journal of Formal Logic*, VII(1):103–105, 1966.

[3] Diderik Batens. A universal logic approach to adaptive logics. *Logica universalis*, 1:221–242, 2007.

[4] Gerhard Brewka. *Nonmonotonic Reasoning: Logical Foundations of Commonsense*. Cambridge University Press, 1991.

[5] Gerhard Brewka, Jurgen Dix, and Kurt Konolige. *Nonmonotonic Reasoning*. CSLI Publications, 1997.

[6] Gerhard Brewka and Thomas Eiter. Prioritizing default logic. In Risto Hilpinen, editor, *Intellectics and Computational Logic: Papers in Honor of Wolfgang Bibel*. Kluwer Academic Publishers, 2000.

[7] Newton C. A. da Costa. On the theory of inconsistent formal systems. *Notre Dame Journal of Formal Logic*, XV(4):497–510, 1974.

[8] Newton C. A. da Costa and E. H. Alves. A sematicall analysis of the calculi c_n. *Notre Dame Journal of Formal Logic*, XVIII(4):621–630, 1977.

[9] John Horty. Defaults with priorities. *Journal of Philosophical Logic*, 36(4):367–413, 2007.

[10] Stanislaw Jaskowski. A propositional calculus for inconsistent deductive systems. *Logic and Logical Philosophy*, 7:35–56, 1999.

[11] Richards J Heuer Jr. *Psychology of Intelligence Analysis*. Nova Biomedical, 1999.

[12] David Makinson. *Bridges from Classical to Nonmonotonic Logic*. College Publications, 2005.

[13] Karl Popper. *The Logic of Scientific Discovery*. Routledge, 1992.

[14] Karl Popper. *Conjectures and Refutations*. Routledge, 2002.

[15] Graham Priest. Minimally inconsistent lp. *Studia Logica: An International Journal for Symbolic Logic*, 50(2):321–331, 1991.

[16] N. Rescher and R. Manor. On inference from inconsistent premises. *Theory and Decision*, pages 179–217, 1970.

[17] Nicholas Rescher and Robert Brandom. *Logic of Inconsistency: A Study in Nonstandard Possible-World Semantics and Ontology*. Blackwell, 1980.

[1] In effect we treat a label as labeling the conjunction of all propositions with that label in the set.

[2] As will be the case e.g. if $\succ$ is a total order.

[3] See, for example, [10, 2, 7, 8, 17, 15, 1, 3].

[4] Brewka[4] and Brewka, Dix and Konolige[5] provide good surveys of non-monotonic logic and Makinson[12] provides an excellent overview of the meta-theory of non-monotonic logic. Whereas Brewka and Eiter [6] and Horty [9] discuss the extension of logics to prioritized systems.

Appendix

TABLE II
The Ordering

DEFINITION > INTERCEPT, DEFINITION > POLICE

INTERCEPT > WEAPONS
INTERCEPT > ATTACKER_TYPE, INTERCEPT > ATTACKER_PREFERENCES, INTERCEPT > SINGLE_ATTACK
INTERCEPT > SINGLE_BOMB_TYPE, INTERCEPT > SINGLE_ATTACK

POLICE > WEAPONS, POLICE > ATTACKER_TYPE, POLICE > ATTACKER_PREFERENCES
POLICE > SINGLE_ATTACK, POLICE > SINGLE_BOMB_TYPE, POLICE > SINGLE_ATTACK

WEAPONS > SIGHTINGS, ATTACKER_TYPE > SIGHTINGS, ATTACKER_PREFERENCES > SIGHTINGS
SINGLE_ATTACK > SIGHTINGS, SINGLE_BOMB_TYPE > SIGHTINGS, SINGLE_ATTACK > SIGHTINGS

SIGHTINGS > INFORMER1, SIGHTINGS > INFORMER2

INFORMER1 > DEFAULTASSUMPTIONS
INFORMER2 > DEFAULTASSUMPTIONS

TABLE III
Assumptions and Hypotheses

DEFINITION : $\mathit{Professional} \equiv \mathit{Career} \lor \mathit{Disposable}$
DEFINITION : $\mathit{BombThreat} \equiv \mathit{PlaceableBomb} \lor \mathit{CarBomb} \lor \mathit{SuicideBomb}$

WEAPONS : $\mathit{QuantityTheftExplosive} \lor \mathit{QuantityPurchasedPrecursors} \supset \mathit{CarBomb}$
WEAPONS : $\mathit{SmuggledRifle} \lor \mathit{PurchasedRifle} \supset Sniper$
WEAPONS : $\mathit{DistanceAttack} \supset \mathit{Mortar}$
WEAPONS : $\mathit{SmallTheftExplosive} \lor \mathit{SmallPurchaseExplosive} \supset \mathit{PlaceableBomb} \lor \mathit{SuicideBomb}$
WEAPONS : $QuantityTheftExplosive \lor \mathit{SmallTheftExplosive} \supset \mathit{Professional}$

ATTACKER_TYPE : $\mathit{Support} \supset \mathit{Professional}$
ATTACKER_TYPE : $\mathit{QuantityPurchasedPrecursors} \supset \mathit{Career}$
ATTACKER_TYPE : $\mathit{EscapePlan} \supset \mathit{Career}$
ATTACKER_TYPE$\neg\mathit{EscapePlan} \supset \mathit{Disposable} \lor \mathit{Amateur}$

ATTACKER_PREFERENCES : $\mathit{Disposable} \supset \mathit{SuicideBomb} \lor \mathit{CarBomb}$
ATTACKER_PREFERENCES : $\mathit{Career} \supset \mathit{PlaceableBomb} \lor \mathit{Mortar}$
ATTACKER_PREFERENCES : $\mathit{Amateur} \supset \mathit{SuicideBomb} \lor \mathit{CarBomb} \lor \mathit{Sniper}$
ATTACKER_PREFERENCES : $\mathit{Mortar} \supset \mathit{Professional}$

SINGLE_ATTACK :
$(\mathit{Career} \land \neg\mathit{Disposable} \land \neg\mathit{Amateur})\lor$
$(\neg\mathit{Career} \land \mathit{Disposable} \land \neg\mathit{Amateur})\lor$
$(\neg\mathit{Career} \land \neg\mathit{Disposable} \land \mathit{Amateur})$
SINGLE_ATTACK :
$(\mathit{Sniper} \land \neg\mathit{Mortar} \land \neg\mathit{BombThreat})\lor$
$(\neg\mathit{Sniper} \land \mathit{Mortar} \land \neg\mathit{BombThreat})\lor$
$(\neg\mathit{Sniper} \land \neg\mathit{Mortar} \land \mathit{BombThreat})$

SINGLE_BOMB_TYPE : $\mathit{BombThreat} \supset$
$(\mathit{PlaceableBomb} \land \neg\mathit{CarBomb} \land \neg\mathit{SuicideBomb})\lor$
$(\neg\mathit{PlaceableBomb} \land \mathit{CarBomb} \land \neg\mathit{SuicideBomb})\lor$
$(\neg\mathit{PlaceableBomb} \land \neg\mathit{CarBomb} \land \mathit{SuicideBomb})$

DEFAULT_ASSUMPTIONS : *Amateur*

HYPOTHESES
Amateur
Professional
- *Career*
- *Disposable*

BombThreat
- *PlaceableBomb*
- *SuicideBomb*
- *CarBomb*

Sniper
Mortar

Networks of Trust and Distrust: Towards Logical Reputation Systems

W. T. Harwood, J. A. Clark, J. L. Jacob

University of York
Department of Computer Science

Abstract. We introduce the notion of a network of trust and distrust relations between individuals and take an argumentation approach to the assessment of whether one individual should trust another.

. . . good decision is based on knowledge and not on numbers”

Plato - Early Dialogues - Laches

1 Introduction

This paper reports ongoing work in creating a logical foundation for reasoning about trust and trustworthiness in networks of individuals that may recommend one another as trustworthy or untrustworthy. One solution is to adopt some form of voting or counting scheme as in commonly done in reputations systems [10]. But in many circumstances, when the stakes are sufficiently high, e.g. deciding to trust a root certificate or disclose confidential information, weight of numbers does not constitute a good argument. As Plato puts it ". . . good decision is based on knowledge and not on numbers"[1].

One of the ultimate goals of this work is to provide the foundations for a logically well founded trust management system, or *Logical Reputation System*, where ‘reputation’ is computed by maintaining some notion of consistency between trust assertions made by trusted individuals. This approach contrasts

[1] For those without a classical education, or more relevantly today, an Internet connection, this is part of a general argument that Plato directs against amalgamating opinions as a basis for reaching a good decision. This is actually a cornerstone of Plato’s arguments against democracy. Today we take a more liberal view and regard some decisions as being appropriately arrived at by amalgamating individual opinions (such as who should rule the country, or what colour should we paint the school) and other decisions as arrived at by knowledge. It is certainly the contention of this paper that trust is best arrived at through knowledge rather than opinion.

with trust models, such as those of Coleman[6] or Marsh[12], that appeal to probabilities of trustworthiness or any similar numeric notions of degree of trustworthiness. Rather, in the approach considered here, a trust judgment is a purely logical resolution of possibly conflicting trust arguments. The intent is to use such a system to automatically make trust judgements in social network applications based on relational information gathered from users. This paper aims at setting out a logical framework based on argumentation theory to achieve this goal.

Our starting point is to consider networks of individuals that assert that they trust some individuals and distrust others.

Trust and distrust[2] are statements about the relationship between two individuals in relation to some action, such as, *information disclosure*, that holds in some context, such as, *today, in this building* (see, for example, Hardin's discussion in [9]). Throughout this paper we will consider the action and context as fixed so that we may talk of trust and distrust as binary relations. It should be apparent that we can put the additional dimensions back into the picture by considering families of relations parameterized by action and context.

If we only had information to the effect that certain individuals were trustworthy we would have a *web of trust* model (see, e.g. Zimmermann [13]) in which one individual trusts another if there is a trusted path between them. Here we consider how such models may be extended in the presence of additional negative assertions to the effect that certain individuals distrust one another. This allows the possibility of a trust path being undermined by a *distrust path*. Here we present a model of such systems in three stages of increasing complexity.

The first stage, *simple trust systems*, captures the idea that an individual trusts another if there is a trust path between them that is not undermined by distrust. Simple trust systems are modeled after argumentation theory[3, 4, 7]. Essentially, the approach is to assess the soundness of the argument that an individual, x_0 say, can trust an individual x_n. In our case the argument for trust is the existence of a trust path between x_0 and x_n in a network of trust relations. However, this argument may be undermined by an attack on it. An attack is an argument that some link in the chain of trust from x_0 to x_n is untrustworthy. In our case, such an argument is the existence of a path of trust from x_0 to some node y_m such that y_m *distrusts* some node connecting x_0 and x_n (including x_n itself). The existence of such an attack would make the original argument unsound, unless, of course, the attack itself was attacked in a similar manner, etc. etc.

[2] The relationship between trust and distrust is far from uncontroversial, see, for example, the discussions in the collection of articles [8]. We take distrust as more than the mere absence of trust. That is, distrust is not simply the complement of trust. Rather, trust and distrust are two relations that can exist between individuals and it is even possible for an individual to trust and distrust another individual simultaneously about the same topic. In such cases, although the individual is conflicted about trust, they are not logically inconsistent about trust.

The argumentation theory approach to resolving the set of attacks and counterattacks is to say that the original argument is sound if it is possible to partition the set, S, containing the original argument and the closure of all the attacks and counter attacks possible based on the initial argument, into two distinct sets, which we call S^+ and S^-, such that: S^+ is consistent in that no paths in S^+ attack one another; S^+ contains the original trust path; and for every path in S^- that attacks a path in S^+, S^+ contains a path that counter attacks that path.

Although formally straightforward, simple trust systems fail to capture an important aspect of trust: that when faced with a choice over conflicting recommendations of who to trust we have preferences over the choices. This leads to the formulation of the second stage, *preferential trust systems*, which introduces the notion that individuals may rank the other individuals into a partial ordering indicating their relative efficacy at making trust or distrust recommendations. This relative ranking is then extended to a partial order on paths which is used to measure the relative strength of paths. A distrust path can only undermine another path if it is sufficiently strong when compared to the path it is attacking (up to the point of attack). This second form of system is formalized by revising the notion of attack between paths.

The final stage *asymmetric preferential trust systems* addresses the fact that, in many situations, individuals have an asymmetric attitude to trust and distrust in that they are more willing to accept an argument that leads them to distrust than they are to accept one that leads them to trust. In the approach considered here, individuals require stronger arguments to make them trust than they do to make them distrust.

In order to directly describe the relationship between individuals, individuals' efficacy assessments, trust paths and distrust paths, trust systems are described relatively concretely. Of course these systems may be considered more abstractly using Dung's abstract argumentation systems framework. The connection between trust systems and Dung's framework is sketched in section 7.

2 Trust Systems

First we set out the framework of trust systems that we use throughout the paper.

A trust system is a collection of individuals I each of which may assert some collection of propositions, P_i for $i \in I$, and two binary relations $Trust : I \Leftrightarrow I$ and $Distrust : I \Leftrightarrow I$. If an individual, say x_0, trusts another individual, say x_n, then x_0 accepts P_n as true. If however x_0 distrusts x_n then x_0 neither accepts P_n as true nor rejects P_0 as false.

Informally, a trust system is a collection of individuals each of which may make assertions about the state of the world. In particular, each individual may assert whether or not they regard some other individuals as trustworthy or untrustworthy. If an individual i regards an individual j as trustworthy we will say that i trusts j. If, on the other hand, i regards j as untrustworthy we will say that i distrusts j. It is also possible for i to neither trust nor distrust j. If i trusts j then i is willing to accept j's assertions as true. In particular i accepts j's assertions about the trustworthiness of others as true. If i accepts a trust assertion of j as true e.g. if j trusts k, then i accepts there is an argument for trusting k, specifically i trusts j and j trusts k. If, however, j distrusts k then i accepts there is an argument that k is untrustworthy i.e. j whose assertions i trusts, distrusts k. It should be clear at this point that trust arguments can be extended (i.e. if i trusts j, j trusts k and k trusts l, then there is an argument for i trusting l) but distrust arguments cannot (i.e. if i trusts j, j distrusts k and k trusts l, then, since j does not accept k's assertions there is neither a trust argument nor a distrust argument, derivable from these facts alone, linking i and l).

Formally a trust system is a collection of individuals I and two binary relations $Trust : I \Leftrightarrow I$ and $Distrust : I \Leftrightarrow I$. Arguments for the trustworthiness and untrustworthiness of individuals will be modeled as trust paths and distrust paths between individuals. A trust path from x_0 to x_n is a sequence $<x_0, x_1, \ldots, x_{n-1}, x_n>$ such that every pair (x_i, x_{i+1}) is in $Trust$. A distrust path from x_0 to x_n is a sequence $<x_0, x_1, \ldots, x_{n-1}, x_n>$ such that every pair (x_i, x_{i+1}) for $i < n-1$ is in $Trust$ and (x_{n-1}, x_n) is in $Distrust$. That is, the path $<x_0, x_1, \ldots, x_{n-1}>$ is a trust path and the final step $<x_{n-1}, x_n>$ is distrusting.

The set of trust paths will be called TP and the set of distrust paths will be called DP.

Given a path, p, (either a trust path or a distrust path) then *range* p is the set of all individuals in the path i.e. if $p = <x_0, \ldots, x_n>$ then $range\ p = \{x_0, \ldots, x_n\}$. We will also say that $first\ p = x_0$ and $last\ p = x_n$, and, for later use, $front\ p = <x_0, \ldots, x_{n-1}>$.

A distrust path, q, *attacks* a path if it attacks the trust supporting the path, meaning it either attacks any point of a trust path (including its last node) or it attacks any point on the front of a distrust path (i.e. the trust path part of the distrust path).

Let $tr\ p$ be the *trust part* of a path p, defined by

$$tr\,p = \begin{cases} p & \textbf{if } p \in TP \\ front\,p & \textbf{if } p \in DP \end{cases}$$

Then *attacks* relation between paths is defined by:

$$q\ attacks\ p \equiv q \in DP \wedge first(q) = first(p) \wedge last(q) \in range(tr\,p)$$

An attack is admissible if it satisfies an *admissibility condition* that varies between the three forms of trust system considered. For simple trust systems all attacks are admissible. For preferential trust systems an attack is admissible only if it is of adequate strength. For asymmetric trust systems the strength condition varies depending on the way the attack will affect the overall outcome. The following section illustrates the effect of the different admissibility conditions with a short example.

3 A Short Example

To illustrate the three systems consider the following network of trust and distrust.

- Alice trusts Bob,
- Bob trusts Carol,
- Carol trusts Dan,
- Dan distrusts Bob,
- Alice trusts Elizabeth,
- Elizabeth distrusts Dan.
- Does Alice trust Carol?

Under simple trust systems, where we have no other information, Dan's distrust of Bob defeats the chain of trust connecting Alice and Carol but Elizabeth's attack on Dan defeats it, and so cancels its effect, leaving Alice having trust in Carol.

If additionally we know:

- Alice rates herself, Bob and Carol strictly higher in ability to make trust judgements than she does Dan.

Then, under a preferential trust system, Alice will trust Carol because Dan's distrust of Bob will lead to an attack path that is weaker than the trust path between Alice and Carol. If, however,

- Alice rates herself, Bob, Carol and Dan equally in ability to make trust judgements

then we would need to consider the exact formulation of the preference system: does an attack from a path of equally strength defeat the attacked path or not? Below we differentiate between *conservative* systems in which attacks must be strictly stronger to defeat a path and *paranoid* systems in which attacks from paths of equal strength, or attacks from incomparable paths, can defeat the attacked path.

Finally, to illustrate asymmetric trust systems, which are *conservative* if the consequence is trust and *paranoid* if the consequence is distrust, we consider two situations

1. Alice rates everyone equally in their ability to make judgements.
2. Alice rates Elizabeth higher than everyone else in her ability to make trust judgements.

Under assumption 1, Dan's distrust of Bob will lead to Alice not trusting Carol even though Elizabeth distrusts Dan, because the distrusting outcome is favored over the trusting outcome. Whereas under assumption 2, Elizabeth's distrust of Bob can cancel the attack and lead to Alice having trust in Carol.

The rest of the paper provides the technical details of each of the systems.

4 Simple Trust Systems

As mentioned above, in simple trust systems all attacks are admissible.

We wish to define notion of a *sound* trust path, p, between two individuals as a path which is either not attacked or only attacked by distrust paths that are themselves defeated by other attacks. To do this we first define the attack closure set of the path p to be p^a , the least set closed under:

- $p \in p^a$,
- $q \in p^a$ and $r \in attacks(q) \implies r \in p^a$.

and say p is sound if and only if p^a can be partitioned into two sets S^+ and S^- such that:

- $p \in S^+$,
- S^+ is *consistent* in that no path in S^+ attacks any other path in S^+, i.e. $S^+ \cap attacks^{\exists}(S^+) = \emptyset$,
- S^+ *defends itself* against S^- in that every path in S^- that attacks a path in S^+ is itself attacked by a path in S^+, i.e. $\overset{\smile}{attacks}{}^{\exists}(S^+) \subseteq attacks^{\exists}(S^+)$.

where, for a relation R, $R^{\exists}X$ is the forward image of X under R i.e. $\{y | \exists x \in S.xRy\}$.

We will call a set, S, a *support*, if it is consistent and defends itself (i.e. it can support some trust path p).

We say an individual x_0 *trusts* an individual x_n iff[3] there is a sound trust path between x_0 and x_n.

Given a simple trust system $T = (I, \{P_i\}_{i \in I}, Trust, Distrust)$ its set of sound trust paths (STP) is defined by:

$$STP = \{(x_0, x_n) \in I \times I \mid \exists p \in TP.first(p) = x_0 \wedge last(p) = x_n \wedge sound(p)\}$$

Two simple trust systems S and T are *trust equivalent* iff they have the same set of individuals and the same set of sound trust paths.

5 Preferential Trust Systems

Preferential trust systems restrict admissible attacks using a notion of relative strength between the attacked path and the attacking path. The particular notion that we use is that the strength of the path is derived from the competence, trustworthiness or reliability of the individuals in the path in making judgments about other individuals. We will settle on the neutral term *efficacy* for any of the terms competence, trustworthiness or reliability (or any other such notion).

In the above, all individuals have been regarded as of equal efficacy in rating the trustworthiness of other individuals. We will now consider what

[3] Here, and throughout, we will adopt the convention of writing *iff* for *if and only if*.

happens when individuals are partially ordered by their efficacy in performing such rating. We will assume every individual i has available their own assessment of the relative efficacy of all other individuals at rating the trustworthiness of others. Formally we take this to be a family of partial orders (reflexive, anti-symmetric and transitive binary relations) over the set of individuals I, one for each member $i \in I$, denoted $\succeq_i$ reflecting i's view of the relative efficacy of individuals. Our goal is that, given a path p which is attacked by a path q, we wish to compare the strength of p up to the point of the attack, $last(q)$, with the strength of q. To do this we need to derive a partial ordering of paths from the partial ordering of the efficacy of the individuals in the paths.

We will call the segment of the path p up to the attack, $p\ |_{last(q)}$. If we were to use a strict total ordering to compare paths then we would say that one path, say q, was weaker than another, say p, when $range(q)$ contained an element less than any element in $range(p)$. We generalize this idea to partial orders by considering minimal elements in the ranges of the paths.

First we define an extension of a partial order over a set to a partial order over subsets of that set.

Given a partial order, $\succeq$, over a set S, we say that a subsets P and Q of S are comparable[4], written $P \sim Q$ iff:

$$(\forall x \in P.\ \exists y \in Q.\ x \succeq y) \vee (\forall y \in Q.\ \exists x \in P.\ x \succeq y)$$

The set of minimal elements of a set P i s defined as:

$$minimal(P) = \{x \in P \mid \forall y \in P.\ (y \succeq x) \implies y = x\}$$

A set, $P \subseteq S$, is at-least-as-strong-as a set, $Q \subseteq S$, written $P \sqsupseteq Q$, iff

$$P \sim Q \wedge \forall x \in minimal(P).\ \exists y \in minimal(Q).\ x \succeq y$$

A set P subset of S is stronger than a set Q subset of S, written $P \sqsupset Q$, iff

$$P \sqsupseteq Q \wedge Q \not\sqsupseteq P$$

All this amounts to is that subsets are ordered by comparing the least elements of the chains and if one of the subsets has strictly smaller elements for any of its chains (and the other does not) then it is the smaller set.

A path p is *stronger than* q, also written $p \sqsupset q$, iff

$$\begin{gathered} first(p) = first(q)\ \wedge last(p) = last(q) \wedge \\ range(p) \setminus \{last(p)\} \sqsupset range(q) \setminus \{last(q)\} \end{gathered}$$

The removal of the last elements of the paths is due to the fact that we derive the efficacy of the individuals on the path that make the trust recommendations.

[4] **Warning**: For those familiar with the notation $x \parallel y$ for x incomparable with y under the partial order $\preceq$. The notion defined here is over subsets of the ordering, not elements of the ordering. So $P \sim Q \equiv \exists p \in P, q \in Q. \neg(p \parallel q)$.

We now modify the definition of attack to take account of the relative strength of paths. There are two possible views of relative strength that correspond to whether the individual x_0 takes a *conservative* or a *paranoid* stance with respect to attacks. If x_0 takes a conservative stance, then a path is only defeated by a strictly stronger attack. If, on the other hand, x_0 takes a paranoid stance, then a path is defeated if the attacking path is incomparable or is at-least-as-strong-as the attacked path. The paranoid position allows attacks to defeat other attacks if x_0 is not in a position to positively assert that the attacked path is the stronger of the two.

That is, if x_0 has a conservative stance, then an attack, q, on a path, p, only succeeds if $q \sqsupset p \mid_{last(q)}$:

$$q\ attacks_C\ p \equiv q \in DP \wedge first(q) = first(p) \wedge last(q) \in range\ p \wedge q \sqsupset p \mid_{last(q)}$$

and if x_0 has a paranoid stance, then an attack, q, on a path, p, only succeeds if $p \mid_{last(q)} \not\sqsupseteq q$:

$$q\ attacks_P\ p \equiv q \in DP \wedge first(q) = first(p) \wedge last(q) \in range\ p \wedge p \mid_{last(q)} \not\sqsupseteq q$$

Preferential trust systems are formulated by replacing the definition of attack in simple trust systems with either the conservative or the paranoid definition of attack[5].

6 Asymmetric Preferential Trust Systems

In practice individuals are often asymmetric in their attitude to trust and distrust. That is, they are paranoid about trust and conservative about distrust. This means that the admissibility of an attack changes according to the overall role it plays in determining the outcome, introducing an asymmetry between paths which ultimately lead to a trust decision and paths which ultimately lead to a distrust decsion. We capture this asymmetry by redefining the conditions for forming S and forming the partitions S^+ and S^-:

- The attack closure set S is the least closed set of paranoid attacks based on a trust path p as above.
- S^+ is restricted to only containing the initial trust path, p, and conservative attacks.
- Since conservative attacks are a subset of paranoid attacks, S^- may contain both types of attack.

Trust path p is sound iff it is possible to form a partition of S such that:

[5] A system may also be formulated where the stance varies from individual to individual which is essentially a simple trust system with an indexed family of *attacks* operators.

- $p \in S^+$.
- S^+ is *consistent* in that no path in S^+ attacks any other path in S^+.
- S^+ is *conservative* in that every path in S^+ is either p or a member of $attacks_C(x)$ for some x.
- S^+ *defends itself* against S^- in that every path in S^- that attacks a path in S^+ is itself attacked by a path in S^+.

7 Connecting Trust and Dung's Abstract Argumentation

Dung [7] defines an abstract argumentation system as a pair $(AR, Attacks)$ where AR is a set of arguments and $Attacks$ is a binary relation over AR called the *attacks* relation. We write $x \ Attacks \ y$ for x attacks y. A set, $S \subseteq D$, attacks an argument, $x \in D$, if some argument in S attacks x (we will say that $S \ ATTACKS \ x$ for $\exists y \in S. \ y \ Attacks \ x$).

Dung then goes on to define the notions of:

- *Conflict free*: A set of arguments $S \subseteq AR$ is *conflict free* iff there is no pair of arguments $x \in S$ and $y \in S$ such that $x \ Attacks \ y$.
- *Acceptable*: An argument $x \in AR$ is *acceptable with respect to S* iff for every argument $y \in AR$ if $y \ Attacks \ x$ then $S \ ATTACKS \ y$. Following [2] we will also say that S *defends* x when x is acceptable with respect to S.
- *Admissible*: A set $S \subseteq AR$ is *admissible* iff S is *conflict free* and each argument in S is *acceptable with respect to S*.

Dung then goes on to discuss various notions of semantics that further restrict the notion of admissibility which are not used in our current semantics.

To translate the above into Dung's framework we consider an individual a and the set of trust paths, P, rooted at a. For the simple trust systems:

- The set of arguments AR is the set P.
- The attacks relation between holds $q, p \in P$, i.e. $q \ Attacks \ p$, iff there exists a distrust path $d = \lt x_0, x_1, \ldots, x_{n-1}, x_n \gt$ with $q = \lt x_0, x_1, \ldots, x_{n-1} \gt$ and $d \ attacks \ p$.

Note that this definition of *Attacks* loses information by conflating multiple distinct attacks from q to different points on p.

A path p is $sound_D$ iff P can be partitioned into two sets S^+ and S^- such that $p \in S^+$ and S^+ is admissible.

Clearly the above notions of consistency and conflict freeness are the same (albeit on different domains):

Proposition 1. $S \cap R^{\exists}(S) = \emptyset \equiv S \subseteq \overline{R^{\exists}(S)}$

Likewise, S defends itself and S is acceptable are essentially the same as demonstrated by the following two propositions.

First we introduce the dual of the forward image operator on binary relations over a set S: Given a binary relation $R : S \leftrightarrow S$, the function $R^{\forall} : \mathcal{P}S \to \mathcal{P}S$ is defined by:

$$R^{\forall}Y = \{x \mid \forall y.xRy \implies y \in Y\}$$

$R^{\exists}X$ is the forward image of X and $R^{\forall}Y$ is the set of elements in the inverse image of R that only result in elements in Y. [6].

$R^{\exists}$ and $R^{\forall}$ form a (covariant) galois connection, or axiality, over S. This means that $R^{\exists} \circ R^{\forall}$ is an interior operator on S and $R^{\forall} \circ R^{\exists}$ is a closure operator on S. Letting $\overset{\smile}{R}$ represent the converse of R (i.e. $x \overset{\smile}{R} y \equiv yRx$) then

Proposition 2. *S is acceptable iff $S \subseteq (\overset{\smile}{Attacks})^{\forall}(Attacks^{\exists}(S))$*

Proof Sketch. This follows from $R^{\forall}X = \overline{(\overset{\smile}{R})^{\exists}(\overline{X})}$ and Amgoud & Cayrol's theorem, quoted in [2], which rendered in our notation is S is *acceptable* iff : S is acceptable iff $S \subseteq \overline{Attacks^{\exists}(\overline{Attacks^{\exists}S})}$.

Proposition 3. *S is acceptable iff $(\overset{\smile}{Attacks})^{\exists}S \subseteq Attacks^{\exists}S$*

Proof Sketch.

$\implies$

$S \subseteq (\overset{\smile}{Attacks})^{\forall}(Attacks^{\exists}(S))$

by proposition 2

$(\overset{\smile}{Attacks})^{\exists}S \subseteq (\overset{\smile}{Attacks})^{\exists}((\overset{\smile}{Attacks})^{\forall}(Attacks^{\exists}(S)))$

by $(\overset{\smile}{Attacks})^{\exists}$ preserves order

$(\overset{\smile}{Attacks})^{\exists}S \subseteq Attacks^{\exists}(S)$

by $(\overset{\smile}{Attacks})^{\exists} \circ (\overset{\smile}{Attacks})^{\forall}$ being an interior operator

$\impliedby$

$(\overset{\smile}{Attacks})^{\forall}((\overset{\smile}{Attacks})^{\exists}S) \subseteq (\overset{\smile}{Attacks})^{\forall}(Attacks^{\exists}S)$

by $(\overset{\smile}{Attacks})^{\forall}$ preserves order

$S \subseteq (\overset{\smile}{Attacks})^{\forall}(Attacks^{\exists}S)$

by $(\overset{\smile}{Attacks})^{\forall} \circ (\overset{\smile}{Attacks})^{\exists}$ being an closure operator

Proposition 4. *$sound_D(p) \equiv sound(p)$*

Proof Sketch. Since the definitions of S being a support and S being admissible are essentially the same between the two definitions of soundness, the major work falls on showing that the existence of a suitable partition of p^a is equivalent to the existence of a suitable partition of P.

[6] $R^{\forall}Y$ is closely related to the weakest precondition operator in programming languages semantics. The exact relation depending on the particular relational theory of programs and termination used.

Recall

$$tr(p) = \begin{cases} p & \textbf{if } p \in TP \\ \mathit{front}(p) & \textbf{if } p \in DP \end{cases}$$

Let p be a trust path, then $tr^{\exists}(p^a) \subseteq P$. Assume the pair S^+, S^- form a suitable partition of p^a then the pair $tr^{\exists}(S^+), P \setminus tr^{\exists}(S^+)$ form a suitable partition of P.

Conversely, if the pair S^+, S^- form a suitable partition of P then the pair $\breve{tr}^{\,\exists}(S^+) \cap p^a, p^a \setminus (\breve{tr}^{\,\exists}(S^+) \cap p^a)$ form a suitable partition of p^a.

To obtain the corresponding Dungian systems for preferential and asymmetric trust systems we modify the definition of the *Attacks* relation. Given the conflating of attacks mentioned above we must ensure that the potential multiplicity of attacks is correctly dealt with when comparing strength.

For two paths $p, q \in P$ such that $q \ \mathit{Attacks} \ p$ we define:

$$q \sqsupset_C p \equiv \forall x \in \mathit{range}(p).(\mathit{last}(q), x) \in \mathit{Distrust} \implies q \sqsupset p|_x$$

$$q \sqsupset_P p \equiv \forall x \in \mathit{range}(p).(\mathit{last}(q), x) \in \mathit{Distrust} \implies p|_x \not\sqsupset q$$

And given a partition of P into S and $\overline{S}$ (= $P \setminus S$):

$$\sqsupset_A^S \equiv ((S \times \overline{S}) \cap \sqsupset_C) \cup ((\overline{S} \times S) \cap \sqsupset_P)$$

Using these orderings we define the three corresponding attacks relations as:

- Conservative Preferential Trust: $\mathit{Attacks}_C = \mathit{Attacks} \ \cap \sqsupset_C$.
- Paranoid Preferential Trust: $\mathit{Attacks}_P = \mathit{Attacks} \ \cap \sqsupset_P$.
- Asymmetric Trust: $\mathit{Attacks}_A^S = \mathit{Attacks} \ \cap \sqsupset_A^S$.

Finally we demonstrate that the asymmetric trust systems have a pleasing simplification of the acceptability condition in that $\mathit{Attacks}_A$ factors into $\mathit{Attacks}_P$ and $\mathit{Attacks}_C$ on either side of the acceptability condition.

Proposition 5. $\mathit{Attacks}_A^S = ((S \times \overline{S}) \cap \mathit{Attacks}_C) \cup ((\overline{S} \times S) \cap \mathit{Attacks}_P)$

Proof Sketch. by boolean algebra

Proposition 6. *An set, S, is acceptable in the asymmetric trust system iff* $(\breve{\mathit{Attacks}}_P)^{\exists} S \subseteq (\mathit{Attacks}_C)^{\exists} S$

Proof Sketch.

$$(\overset{\smile}{Attacks^S_P})^{\exists} S \subseteq Attacks^{S\,\exists}_C S$$

by proposition 2

$$(\overset{\smile}{((S \times \overline{S}) \cap Attacks_C) \cup ((\overline{S} \times S) \cap Attacks_P)}\ ^{\exists})S \subseteq$$
$$(((S \times \overline{S}) \cap Attacks_C) \cup ((\overline{S} \times S) \cap Attacks_P)^{\exists})S$$

by proposition 5

$$\overset{\smile}{((S \times \overline{S}) \cap Attacks_C)}\ ^{\exists}S \cup \overset{\smile}{((\overline{S} \times S) \cap Attacks_P)}\ ^{\exists}S \subseteq$$
$$((S \times \overline{S}) \cap Attacks_C)^{\exists}S \cup ((\overline{S} \times S) \cap Attacks_P)^{\exists}S$$

by distribute $(\_)^{\exists}$ over union

$$(\overset{\smile}{Attacks_P})^{\exists}S \subseteq (Attacks_C)^{\exists}$$

by domain restrictions

8 Conclusions

For us at least, the idea of using argumentation to reason about networks of trust, and distrust, is in its infancy. The work presented here raises more questions than it answers, some of which we raise below (and there are many more than raised here).

Trust systems as outlined above offer a logically well founded approach to reasoning about trust based on minimal information gathered from individuals i.e. the individuals relative assessment of the efficacy of the judgements of others and a map of immediate trust and distrust relations between individuals. The natural next step is to investigate this in practice in an actual social network application.

The asymmetric preferential trust systems above rely on the fact that conservative attacks are a subset of paranoid attacks. Clearly it is possible to generalize further and define relevant attacks and acceptable rebuttals to relevant attacks. Given a set of attacks, we classify some attacks as relevant, some as acceptable rebuttals of relevant attacks, and some as neither. S is built as the closure of attacks on a trust path p as above and we define S^+ and S^- by:

- $p \in S^+$,
- S^+ is *consistent* in that no path in S^+ attacks any other path in S^+,
- S^+ is a *rebuttal set* in that every path in S^+ is either p or a rebuttal attack,
- S^+ *defends against relevant attacks* from S^- in that every relevant attack in S^- that attacks a path in S^+ is itself attacked by a path in S^+.

This generalization opens up the possibility of considering richer asymmetries between trust and distrust arguments. For example, if we drop the use of the extended order relation and consider using a labeling of the individuals in paths. Consider, as illustration, a sensor network based on three kinds of individual sensor: electronic sensing and people that perform either casual or detailed inspections. We may trust an individual because we have a mixed trust path to it but relevant attacks may be limited to chains that exclude electronic sensors and rebuttals may be limited to chains of people who perform detailed inspection[7]. This approach will be the subject of further investigation.

The relation to Dungian argumentation outlined in section 7 uses only the most basic semantic notion of admissibility. This raises the question whether or not the other possible semantics have a useful meaning for trust (and distrust) relations. The question is why we would *want* a richer set of arguments than that required to support the sounds of a particular trust path? Perhaps there is a useful notion of sets of individuals you can consistently trust corresponding to the other possible semantics. It certainly is worth investigating.

During the revision of this paper the authors encountered the work of Cayrol and Lagasquie-Schiex on Bipolar Argumentation [5] systems, and of Kaci and Torre , and Amgoud, Dimopoulos and Moraitis Preference Based Argumentation (se e.g. [11] and [1] respectively). Both seem to overlap on the intent pursued here and offer interesting directions for future investigation.

9 Reviewer Acknowledgements

The authors would like to thank the reviewers for their constructive criticism, knowledgeable comments and insightful questions. In particular we would like to thank two of the reviews for their detailed comments and references to the work of other authors. This latter has been particularly useful to the ongoing work, even though not adequately reflected in this paper. In addition to minor corrections of the text the comments have been addressed with additional footnotes and the addition of Section 7 (in response to the reviewer plea for more mathematics).

10 Sponsorship Acknowledgements

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

[7] Admittedly, this example can be done using order relations, but it is seems conceptional simpler as a predicate on the acceptable sets of attacks and counter attacks.

Bibliography

[1] Leila Amgoud, Yannis Dimopoulos, and Pavlos Moraitis. Making decisions through preference-based argumentation. In Gerhard Brewka and Jérôme Lang, editors, *KR*, pages 113–123. AAAI Press, 2008.

[2] Philippe Besnard and Sylvie Doutre. Characterization of semantics for argument systems. In Didier Dubois, Christopher A. Welty, and Mary-Anne Williams, editors, *KR*, pages 183–193. AAAI Press, 2004.

[3] Philippe Besnard and Anthony Hunter. *Elements of Argumentation*. MIT Press, 2008.

[4] A. Bondarenko, P.M. Dung, R.A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93:63–101, 1997.

[5] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In Lluis Godo, editor, *ECSQARU*, volume 3571 of *Lecture Notes in Computer Science*, pages 378–389. Springer, 2005.

[6] James Coleman. *Foundations of Social theory*. Belknap Press, 1990.

[7] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

[8] Russell Hardin, editor. *Trust & Distrust*. Russell Sage Foundation, 2004.

[9] Russell Hardin. *Trust*. Polity Press, 2006.

[10] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.

[11] Souhila Kaci and Leendert van der Torre. Preference-based argumentation: Arguments supporting multiple values. *Int. J. Approx. Reasoning*, 48(3):730–751, 2008.

[12] Stephen Paul Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, Dept. of Computing and Mathematics, University of Stirling, 1994.

[13] Phil Zimmermann. *PGP User Guide*. MIT Press, 1994.

Conceptualizing Internet Trust

W. T. Harwood*, J. A. Clark, J. L. Jacob

Will.Harwood@cs.york.ac.uk, jac@cs.york.ac.uk, Jeremy.Jacob@cs.york.ac.uk

University of York
Department of Computer Science

Abstract. We set out to address what kind of theoretical framework is required to discuss Internet Trust if 'trust' is to have the same meaning as it does in other social and business contexts. And, given such a framework, what are its implications for the technical, social and legal mechanisms that must be provided to support that notion of trust? We suggest that the elements of the analysis of trust offered by Luhmann, Barber, Miztal and Giddens provide a framework for discussing trust and that the technical mechanisms for ensuring Internet Trust should be assessed against it. Moreover we claim that assessing current Internet Trust models against such a framework shows that they fail to offer meaningful support for trust. We attempt to illustrate this with two brief examples.

1 Introduction

Luhmann[1] and Barber[2] , independently, propose the role of trust is to create certainty in the face of uncertainty and thus allow action as opposed to endless indecision.

Misztal[3] elaborates this line of thought by discussing the role of different forms of trust, *Habitus*, *Passion* and *Policy*, in creating different forms of order in society that act in concert to reduce uncertainty[4]. In modern societies we go beyond interaction with known individuals and extend trust to *anonymous others*. What reduces the uncertainty in dealing with anonymous others?

Luhmann argues that these interactions in modern societies give rise to abstract mechanisms of interaction, such as Money and Power. These abstract mechanisms, which Luhmann calls *generalized communication media*, substitute

[1] Niklas Luhmann. Trust and Power. John Wiley & Sons, 1979

[2] Bernard Barber. The Logic and Limits of Trust. Rutgers University Press, 1983.

[3] Barabra A. Misztal. Trust in Modern Societies. Polity Press, 1996.

[4] *Habitus* is a system of dispositions acquired by an agent through participation in society and may be seen as the set of implicit *rules* or *conventions* that exists below conscious and rational thought. *Passion* is trust arising from the development of the intimate social bonds between people based on shared values and experiences. It is the "internalized trust" that arises from those basic face-to-face interactions. *Policy* operates at the conscious dispositional level. It is built on the set of dispositions that we consciously hold towards individuals and groups.

for the personal relationships built up in face-to-face societies and their role is to provide *equivalent certainty* to personal trust. But this then leads to a need for *systems trust*, which is the belief in the ability of these abstract mechanisms to guarantee 'equivalent certainty'. Thus, Luhmann sees systems trust as the cost, and necessary underpinning, of the modern, networked, society.

This theme is picked up by Giddens[5]: *"Trust in abstract systems is the condition of time-space distanciation and of the large areas of security in day-to-day life which modern institutions offer compared to the traditional world."*

It is possible to 'trust' anonymous individuals because interaction with them is through generalized communication media, through generalized systems of exchange, that are themselves supported by systems trust. *We posit the general goal of any online trust framework is either, to reflect the societal trust frameworks built on Habitus, Passion and Policy, or to establish a new generalized communication medium to carry the burden of trust. In either case it must be supported by systems trust underwriting the particular mechanism.* Below we discuss how common Internet solutions fail to meet these ends.

2 Some Simple (Counter)Examples

Reflecting Societal Trust - Reputation and Recommender Systems: In many cases when we lack direct experience we make rational judgments about people and things by using our social network. We know someone, who knows someone, who knows X. We have a chain connecting us to X, and we have information about the nodes in that chain that allows us to transfer opinions back to ourselves. That process may be by receiving information along the chain, or it might be by using the chain to create introductions so that we may have direct contact with some of the individuals in the chain. Indeed there may be many such chains and we may amalgamate information gained by exploiting many of them. In amalgamating information we make judgements about the *competence* of the individuals involved in arriving at assessments, the *similarity* of their circumstance to our own and the *authenticity* of an individual's statements (i.e. the likelihood of them making authentic statements about the matter at hand rather than pursuing some covert agenda). This social network is not simply a background *givens*, its existence, its shape and its utility comes about by *use*, *participation* and *reciprocation*. It is not merely that we return the favors derived from the network but that feedback and reinforcement shapes our network.

If we compare our network of social trust to online reputation or recommender systems we quickly see that many elements are missing. There is little opportunity to gain the information to determine the competence, similarity or authenticity of individuals involved in the network. As such we are ill placed to amalgamate their views to arrive at a rational judgement. Moreover shaping our network by participation is often not an option.

[5] Giddens. The consequences of Modernity. Polity Press, 1990.

Creating Equivalent Trust - Online Purchase: Technical infrastructure offers to protect us when we undertake online financial transactions. But that protection fails to amount to providing a generalized medium that offers equivalent certainty. The general pattern for such a medium is "X offers equivalent certainty to Y because of the underlying systems trust in Z". In this case we take Y to be making a purchase in a (real) shop with a credit card, X to be making an online purchase from an online shop with a credit card and we explore the nature of Z. The certainty generate by Y in this case arises from, amongst other things:

- The assurance of physical premises (the existence of premises, the nature of the premises, the location of the premises, the time the business has occupied the premises).
- The applicable law due to location.
- The assurance of staff (their presence, their attitude, their knowledgeability, their competence).
- The attribution of responsibility for the transaction to one or more staff members.
- The contact between your direct social network and the shop and its staff (which includes knowing people who know people that work there etc.).
- The assurance of an exchange transaction i.e. goods in hand (at least in many cases).
- The assurance derived from the social understanding and familiarity of shops.

The technical infrastructure supporting 'safe' online shopping gives no real assurances about the identity of the organization we are dealing with, its permanence, its real location, applicable law and legal domain or the responsible individuals associated with an order. Corresponding assessments to those that we derive from the physical premises and from the interaction with staff are denied us, as are assessments derived from location and our wider social network.

3 Conclusion

Unlike Fahrenholtz and Bartelt[6] , who conclude that true trust on the Internet must always be elusive as it is deeply dependent on social and psychological factors, we see things more positively. Trust can be achieved but to do so we must meet its preconditions. We can reflect societal trust networks in the Internet but only by reflecting the societal processes that give rise to them. We can construct equivalent certainty if we construct mechanisms that provide specific certainties (e.g. legal domain) that are underpinned by abstract system in which we can trust.

[6] Dietrich Fahrenholtz and Andreas Bartelt. Towards a sociological view of trust in computer science. Eighth Research Symposium on Emerging Electronic Markets (RSEEM), 10 2001.

4

4 Acknowledgements

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

H. Full Program Listing

H.1. Overview

The program PCC (Preferential Coherence Calculator) is a small test bed for various versions of preferential coherence. Its basic structure is a pipeline that reads a file from standard in, parses it, calculates a preferential theory according to rules determined by the input format and pretty prints an output on standard out.

A preferential theory is a labelled propositional theory with a binary relation imposed between some of the labels. The binary relation is taken to be a generalised notion of a preference ordering over the labels. However, in general, no particular order conditions are required. If the relation is specified using the key word ‘simple‘ then it is a free relation i.e. the relation is exactly what is specified. If the keyword ‘transitive‘ is used then the transitive closure of the relation is generated. The relation is then used induce an ordering on sets of labels as described in the thesis.

Other kinds of relation are also supported by the program and correspond to other possible approaches to coherence. The key words ′upper′, ′lower′ and ′combined′ correspond to taking the given ordering, making the assumption that it specifies a pre-order, and generating the ordering for the upper, lower and combined power domain.

Use of the key word ‘majority’ disallows the use of an order relation but causes conflicts to be resolved by majority voting. In this case the coalition theory consists of the propositions most supported by agents i.e. if some agent supports p and some agent supports $\neg\, p$ then the coalition theory contains the which ever of the two propositions is entailed by most agents.

The key word ‘majority’ offers a variant on this in which agents may be given a weight, or strength, which they lend to each of their propositions. The coalition theory is arrived at taking the theory in which the weight of p or $\neg\, p$ is used to determine which is included in the theory.

The workhorse of the program is the signed tableaux routine that is used both finds inconsistent sets of propositions and to compute entailments (i.e. it finds proofs that A follows from B). This routine corresponds to the non-deterministic model finder described in chapter 3.

File	Description
AgentLogic.y	YACC style grammar file for parser and lexer function
AgentLogic.hs	Generated Parser and lexer – this file is automatically generated from AgentLogic.y
Sets.hs	Functions for treating lists as sets
Relations.hs	Functions for treating lists of pairs as binary relations
SyntacticRelations.hs	Functions for turning relations defined in the syntax to lists of pairs
Coalitions.hs	Functions for building coalitions i.e. sets of sources. Mainly concerned with extending element orderings to set orderings.
TrustExpansion.hs	Functionality for an experiment with a notion of trust as "speaking as/for another". Not currently relevant.
Tableaux.hs	An implementation of signed tableaux based on Smullyan's book and some wrappers for the functionality.
Beliefs.hs	Contains the function "beliefs" which calculates the alternative preferential theories.
Pretty.hs	A set of pretty print routines for the output uses PPrint library module written by Daan Leijen
Inspect.hs	Mainly a place holder left after much code has been replaced/relocated.
main.hs	Top level driver for read/process/print.

H.2. Listing

```
{
module AgentLogic  where

import Data.Char
import Data.List

}

%name system System
%name orderings Orders
%name ordering OrdSeq
%name agents AgentAssertions
%name sources SourceAssertions
%name formula Formula

%tokentype { Token }
%error { parseError }

%token
        simple          { TokenSimple }
        transitive      { TokenTransitive }
        majority        { TokenMajority }
        weighted        { TokenWeighted }
        upper           { TokenUpper }
        lower           { TokenLower }
        combined        { TokenCombined }
        int             { TokenInt $$ }
        hypothesis      { TokenHypothesis }
        '<=>'           { TokenEquiv }
        '=>'            { TokenImplies }
        '+'             { TokenOr }
        '&'             { TokenAnd }
        name            { TokenName $$ }
        '~'             { TokenNeg }
        '('             { TokenBra }
        ')'             { TokenKet }
        ','             { TokenSep }
{-      '@'             { TokenAt }             -}
        '{'             { TokenCurlyBra }
        '}'             { TokenCurlyKet }
        '='             { TokenEqual }
        '>'             { TokenGreater }
{-      ';'             { TokenSemi }           -}
        ':'             { TokenColon }
        '|'             { TokenBar }




%left '@'
%left '<=>'
%right '=>'
%right '+'
%right '&'
%left "~"
%left NEG


%%

OrderTypes :              simple Orders                 { Simple $2 }
                        | transitive Orders             { Transitive $2 }
                        | majority                      { Majority }
                        | weighted WeightSeq            { Weighted $2 }
                        | upper Orders                  { Upper $2 }
                        | lower Orders                  { Lower $2 }
                        | combined Orders               { Combined $2 }

WeightSeq :               WeightAssign                  { [ $1 ] }
                        | WeightSeq ',' WeightAssign    { $1 ++ [ $3] }

WeightAssign :          QualifiedName ':' int           { ($1, $3) }


Orders :                {- EMPTY -}                     { [] }
                {-      | OrdSeq ',' Orders             {  [$1] ++ $3 }  -}
                        | OrdSeq                                { [ $1 ] }
                        | Orders ',' OrdSeq             { $1 ++ [$3] }

OrdSeq :                  QualifiedName                 {  [GTag $1] }
                        | OrdSeq '>' QualifiedName      { $1 ++ [GTag $3] }
                        | OrdSeq '=' QualifiedName      { $1 ++ [ETag $3] }

System :                OrderTypes '|' AgentAssertions          { ($1, $3, []) }
                        | OrderTypes '|' AgentAssertions        hypothesis Formulalist  { ($1, $3, $5) }

AgentAssertion :        QualifiedName ':' Formula { Says $1  $3  }

AgentAssertions :       AgentAssertion  { [$1] }
                        | AgentAssertions ',' AgentAssertion { $1 ++ [$3] }

SourceAssertion :       '{' Qualifiednames '}' ':' Formula { Source $2 $5 }

SourceAssertions :      SourceAssertion { [$1] }
                        | SourceAssertions ',' SourceAssertion  { $1 ++ [$3] }
```

```
Formulalist :           Formula                                 { [$1] }
                        | Formulalist ',' Formula               { $1 ++ [$3] }

Formula :        Formula '<=>' Formula { Equiv $1 $3 }
                | Formula '=>' Formula { Implies $1 $3 }
                | Formula '+' Formula { Or $1 $3 }
                | Formula '&' Formula { And $1 $3 }
                | '(' Formula ')' { $2 }
                | '~' Formula %prec NEG { Neg $2}
                | QualifiedName { Prop $1 }



Params :        name                    { [$1] }
                | Params ',' name       {  $1 ++ [$3] }

QualifiedName : name                    { Name $1 [] }
                | name '(' Params ')'   { Name $1 $3 }

Qualifiednames : QualifiedName                          { [$1] }
                 | Qualifiednames ',' QualifiedName     {  $1 ++ [$3] }




{

{- change TokenProp to TokenName and prop to name -}

parseError :: [Token] -> a
parseError _ = error "Parse error"

data Formula  =         Prop Names |
                        Neg Formula |
                        And Formula  Formula |
                        Or Formula Formula  |
                        Implies Formula  Formula  |
                        Equiv Formula  Formula
                        deriving (Show, Read, Eq, Ord)


data AgentSays =        Says Names Formula
                        deriving (Show, Read, Eq, Ord)

data Sources =          Source [Names] Formula
                        deriving (Show, Read, Eq, Ord)

data Names =            Name String [String]
                        deriving (Show, Read, Eq, Ord)

data OTags =            GTag Names | ETag Names
                        deriving (Show, Read, Eq, Ord)


data Orders a =         OGreater a a | OEqual a a
                        deriving (Show, Read, Eq, Ord)

data OrderTypes =         Simple [[ OTags ]]
                        | Transitive [[ OTags ]]
                        | Majority
                        | Weighted [ (Names, Int) ]
                        | Upper [[ OTags ]]
                        | Lower [[ OTags ]]
                        | Combined [[ OTags ]]
                        deriving (Show, Read, Eq, Ord)


restructure l =
  let   restruct' s [] = s
        restruct' s ((GTag x):(GTag y):l)       = if l /= [] then  restruct' ((OGreater x y):s)  ((GTag y):l) else ((OGreater x y):s)
        restruct' s ((GTag x):(ETag y):l)       = if l /= [] then  restruct' ((OEqual x y):s)     ((ETag y):l) else ((OEqual x y):s)
        restruct' s ((ETag x):(GTag y):l)       = if l /= [] then  restruct' ((OGreater x y):s)  ((GTag y):l) else ((OGreater x y):s)
        restruct' s ((ETag x):(ETag y):l)       = if l /= [] then  restruct' ((OEqual x y):s)     ((ETag y):l) else  ((OEqual x y):s)
   in
        restruct' [] l

data Token =
                          TokenSimple
                        | TokenTransitive
                        | TokenMajority
                        | TokenWeighted
                        | TokenUpper
                        | TokenLower
                        | TokenCombined
                        | TokenEquiv
                        | TokenImplies
                        | TokenOr
                        | TokenAnd
                        | TokenName String
                        | TokenNeg
                        | TokenBra
                        | TokenKet
                        | TokenSep
                        | TokenAt
                        | TokenCurlyBra
                        | TokenCurlyKet
                        | TokenEqual
                        | TokenGreater
                        | TokenSemi
                        | TokenColon
                        | TokenBar
                        | TokenInt Int
                        | TokenHypothesis
                        deriving Show


lexer :: String -> [Token]
lexer [] = []
lexer (c:cs)
        | isSpace c = lexer cs
        | isAlpha c = lexName (c:cs)
        | isDigit c = lexNum (c:cs)
lexer ('<':'=':'>':cs) = TokenEquiv : lexer cs
lexer ('=':'>':cs) = TokenImplies : lexer cs
lexer ('+':cs) = TokenOr : lexer cs
lexer ('&':cs) = TokenAnd : lexer cs
lexer ('~':cs) = TokenNeg : lexer cs
lexer ('(':cs) = TokenBra : lexer cs
lexer (')':cs) = TokenKet : lexer cs
lexer (',':cs) = TokenSep : lexer cs
lexer ('@':cs) = TokenAt : lexer cs
lexer (':':cs) = TokenColon : lexer cs   {- alternative name for @ -}
lexer ('{':cs) = TokenCurlyBra : lexer cs
lexer ('}':cs) = TokenCurlyKet : lexer cs
lexer ('=':cs) = TokenEqual : lexer cs
lexer ('>':cs) = TokenGreater : lexer cs
lexer (';':cs) = TokenSemi : lexer cs
lexer ('|':cs) = TokenBar : lexer cs
lexer (_:cs) = lexer cs


lexNum cs = TokenInt (read num) : lexer rest
                where (num,rest) = span isDigit cs

lexName cs =
        case span isAlphaNum cs of
                ("simple",rest) -> TokenSimple : lexer rest
                ("transitive",rest) -> TokenTransitive : lexer rest
                ("majority",rest) -> TokenMajority : lexer rest
                ("weighted",rest) -> TokenWeighted : lexer rest
                ("upper",rest) -> TokenUpper : lexer rest
                ("lower",rest) -> TokenLower : lexer rest
                ("combined",rest) -> TokenCombined : lexer rest
                ("hypothesis",rest) -> TokenHypothesis : lexer rest
                (var,rest) -> TokenName var : lexer rest

pa = print . formula . lexer


parseFormula = formula  . lexer

parseAgents = agents . lexer

parseSources = sources . lexer

parseSystem = system . lexer

}
```

```
module Sets where

import Data.List

subset xs ys = and [x `elem` ys | x <- xs ]

equalSets xs ys = and ([x `elem` ys | x <- xs ] ++ [y `elem` xs | y <- ys])


kern [] = []
kern (x:l) = x: (kern [ y | y <- l , not(equalSets x y)])


subsets_down l =
 let
      ll = nub l
      subsets' acc ss =
        let next = [ y | y <- nub [ delete u s | u <- ll, s <-ss ], y `notElem` acc ]
          in
                if (any null ss)  then   acc else (subsets' (acc ++ next) next)
  in
    (subsets'  [ ll ] [ ll ])

notMember x xs = not(any (equalSets x) xs)

subsets_up l =
 let
      ll = nub l
      subsets' acc ss =
        let next = kern [ y | y <- [ u:s | u <- ll, s <-ss ], notMember y  acc ]
          in
                if length (head ss) == length ll  then   acc else (subsets' (acc ++ next) next)
  in
    (subsets'  [ [] ] [ [] ])

remove x [] = []
remove x (y:ys) | x == y = remove x ys
                | otherwise = y:(remove x ys)

difference xs []  = nub xs
difference xs (y:ys)  = difference (remove y xs) ys

bigCap [] = []
bigCap [x] = x
bigCap (x:xs) = x `intersect` (bigCap xs)

bigCup [] = []
bigCup (x:l) = x `union` (bigCup l)

first p [] = []
first p (a @ (x:l)) = if p x then a else first p l
```

```
module Relations where

import Data.List
import Sets
import AgentLogic

close step l =
         let close last l = if equalSets last l then nub l else close l  (l ++ (step l))
         in close [] l

stepTrans l = nub [ (x,z) | (x,yr) <- l, (yl,z) <- l, yr == yl ]



stepSym l = nub [ (y,x) | (x,y) <- l]

stepRef l = nub ([ (x,x) | (x,_) <- l] ++ [ (y,y) | (_,y) <- l])

stepRST l = (stepRef l) `union` ( stepSym l) `union` ( stepTrans l)



closeTrans l = close stepTrans l

closeRST l = close stepRST l

field ord = nub( [ a | (a, _) <- ord] ++ [b | (_,b) <- ord])

least xs = nub [ x | (_, x) <- xs, [y | (x1,y) <- xs, x == x1] == [] ]

bottom elm ord = (least ord) ++ ( elm `difference` (field ord))

{- Useful to wrap a Relations with its field, particularly orderings hence the name -}

{- new Ord compose of field equivalence relation and order relation -}

data OrderedSet a = Ord [a] [(a,a)] [ (a,a)]
                    deriving (Show, Read, Eq, Ord)

ends (Ord elms _ ord) = bottom elms ord

ords (Ord _ _ ord) = ord
equiv(Ord _ equ _ ) = equ
elems (Ord elms _ _) = elms
```

```
module SyntacticRelations where

import Data.List
import AgentLogic
import Relations


extractOEq l = [(x,y) | OEqual x y <- l]
extractOGr l = [(x,y) | OGreater x y <- l]


expand eq gr =
        let expand' eq gr = nub([ (x,y) | (x,y1) <- gr, (y2,y) <- eq,  y1 == y2] ++
                                [ (x,y) | (x1,y) <- gr, (x2,x) <- eq,  x1 == x2] ++ gr)
        in
          expand'  (closeRST eq) gr

po eq gr =  closeTrans (expand eq gr)

{- functions for syntactically defined orders -}


{- transitive ordering -}

mkpo ord =
        let
            eq = extractOEq ord
            gr = extractOGr ord
        in
                po eq gr

mkOrd order = mkpo (restructure  order)
mkOrds order = mkpo (concat (map restructure order))

{- succ = immediate Successor -}

successor ord =
        let
            eq = extractOEq ord
            gr = extractOGr ord
        in
                (expand eq gr)


mkSuccOrd order = successor (restructure  order)
mkSuccOrds order = successor (concat (map restructure  order))
```

```
module Coalitions where

import Data.List
import Sets
import AgentLogic
import Tableaux
import Relations
import SyntacticRelations
import TrustExpansion


{- Ways of lifting a simple ordering to a list ordering -}

{- for all y in ys there exists an x in xs such that x > y -}

upper ord  xs ys =  and [or  [ ((x,y) `elem` ord)  | x <- xs] | y <- ys]


lower ord xs ys =  and [ or [ (x,y) `elem` ord | y <- ys] | x <- xs]

combined ord xs ys  = (upper ord xs ys) && (lower ord xs ys)

image ord x = [y | (x1,y) <- ord, x1 == x]

invImage ord y = [x | (x, y1) <- ord, y1 == y ]

converse ord = [(y,x) | (x, y ) <- ord]

{- invImage r x = image (converse r) x -}

forward ord ys = bigCup (map (image ord) ys)

previous ord xs = bigCup (map (invImage ord) xs)

{- Weak set filters -}


smaller ord xs ys =
        let common = xs `intersect` ys
         in
                (upper (closeTrans ord) (difference ys common) (difference xs common))

smallerX ord xs ys =
        let common = xs `intersect` ys
         in
                (upper (closeTrans (ords ord)) (difference ys common) (difference xs common))

uppereq equ ord  xs ys =  and [or  [ or [((x,y) `elem` ord) , ((x,y) `elem` equ) ] | x <- xs] | y <- ys]

{- uppereq equ ord  xs ys =  and [or  [ ((x,y) `elem` ord)  | x <- xs] | y <- ys] -}

biggereq ord xs ys =  (uppereq (equiv ord)  (ords ord) xs ys)


compbiggereq ord xs ys = biggereq ord (difference (elems ord) ys) (difference (elems ord) xs)



bigger ord xs ys = (biggereq ord xs ys)  && (not (biggereq ord ys xs))

compbigger ord xs ys = bigger ord (difference (elems ord) ys) (difference (elems ord) xs)



filterWeakSetsOld ord l = [ x | x <- l, and[ not(smaller ord x y) | y <- l, x /= y]]

filterWeakSets ord l = [ x | x <- l, not (or[(compbigger ord y x) | y <- l, x /= y]) ]

implies  x y = (not x) || y




defended ord   s =
        let r = ords ord
        in
          and[ or[ ((x1,y) `elem` r) |  x1 <- s]  |  (y,x) <- r, x `elem` s, not(y `elem` s) ]


{- experimemtal covers consistent complement -}

{- filterDefendedSets ord l = [ x | x <- l, biggereq ord x (difference (bigCup l) x)  ] -}

filterDefendedSets ord l = [ x | x <- l, biggereq ord x (difference (elems ord) x)  ]





{- important bits from theory -}

maxcon cons s  =
  let  maxcon' acc ys  =
               case (first cons ys) of
                  x:xs ->  ( maxcon' (x:acc) [y | y <- xs, not(any (subset y) (x:acc)) ] )
```

```
                    [] -> acc
  in maxcon' [] (subsets_down s)

maxSet ord  cons s = bigCap (filterWeakSetsOld ord (maxcon cons s))


consistentFormulaSet fs =  (prove [map L fs]) /= []


consistentAA as agentSet = consistentFormulaSet [ f | a <- agentSet, Says b f <- as , a == b]

maximalSets as ord = filterDefendedSets ord (filterWeakSets ord (maxcon (consistentAA as) (nub [a | Says a _ <- as])) )



{- functions for weighted orderings -}

replace [] x = 0
replace ((a,b):l) x      | x == a = b
                         | otherwise = replace l x

weight weights xs = (sum(map (replace weights) xs))

weighted weights xs ys =
        let wxs = (weight weights xs)
            wys = (weight weights ys)
        in
                if wxs > wys then GT
                else if wxs == wys then EQ
                else LT


{- functions unchanged -}




collectAssertions as = [ f | Says _ f <- expandAgents as]

{- filtering out inconsistent agents from coalitions L-}

collectSupport as = [ Says a f | a <- nub[ a | Says a _ <- as],
                          f <- collectAssertions as,
                         let agentsTheory = [ form | Says x form <- expandAgents as, x == a]
                          in if consistentFormulaSet agentsTheory
                                 then (entails [ form | Says x form <- expandAgents as, x == a] f)
                                 else False ]


collectSupporters as = nub [ (nub [a | Says a f1 <- as, f1 == f], f) | Says _ f <- as]

{-
extractEq l = [ (x,y) | (x,y) <-l, x == y ]

extractGr l = [ (x,y) | (x,y) <- l, x /= y]

{- functions to do with ordrings changed -}

-}


{- ------------------------------ -}


comparisons cmp equ l = nub(    [ OGreater x y | x <- l, y<- l,  cmp x y] ++
                                [ OEqual x y | x <- l,   y<- l,  equ  x y] )


coalitions cmp equ l = comparisons cmp equ (kern(map fst l))

equSets a b = (a /= b) && (equalSets a b)

mkSuccOrd' cmp l = successor(coalitions cmp equSets l)

{- ------------------------------ -}

consistentSource as sourceSet = consistentFormulaSet [ f | a <- sourceSet,  (b, f) <- as , equalSets a b]


maxiSource ord as s = maxSet ord (consistentSource as) s
extenSource as xs s = if  consistentSource as (s:xs) then (s:xs) else xs

consist' as ord x = extenSource as (consistStar' as ord [ y | (y,x1) <- ord, x1 == x ]) x
consistStar' as ord s = maxiSource ord as (bigCup [consist' as ord e | e <- s])

majority xs ys = (length xs) > (length ys)

weigh w xs ys = (weight w xs) > (weight w ys)
```

```
module TrustExpansion where

import AgentLogic
import Tableaux

asserts a as  = [ f  | (Says x f) <- as, x == a]

anything = Prop ( Name "$$" [] )
notAnything = Neg anything
tt = Or anything notAnything
ff = And anything notAnything

foldrOpt f v [] = v
foldrOpt f v l = foldr1 f l

mkAnd  l = foldrOpt And tt  l
mkOr l =   foldrOpt Or ff l

mkTrust as a = mkAnd (asserts a as)
mkDistrust as a = mkAnd (map Neg (asserts a as))
mkWeakDistrust as a = Neg (mkTrust as a)



substProp   f n (a @ (Prop (Name y l)))  = if n == y then f l else a
substProp   f n (Neg y)  = Neg (substProp f n y )
substProp   f n (And  x y) = And (substProp f n x ) (substProp f n y )
substProp   f n (Or  x y) = Or (substProp f n x ) (substProp f n y )
substProp   f n (Implies  x y) = Implies (substProp f n x ) (substProp f n y )
substProp   f n (Equiv  x y) = Equiv (substProp f n x ) (substProp f n y )

{- Tp Do: make params qualified names - the following is a temporary fix for simple names -}


expTrust as =  (\l -> mkAnd(map ((mkTrust as) . (\ x -> Name x [])) l))
expDistrust as = (\l -> mkAnd(map ((mkDistrust as) . (\ x -> Name x [])) l))
expWeakDistrust as = (\l -> mkAnd(map ((mkWeakDistrust as) . (\ x -> Name x [])) l))

expandTrust as f = substProp (expTrust as) "Trusts"  f
expandDistrust as f = substProp (expDistrust as) "Distrusts"  f
expandWeakDistrust as f = substProp (expWeakDistrust as) "WDistrusts"  f

{- To Do: expand agent assertions for Trusts, Distrusts and WDistrusts -}

containsProp pl  (Prop (Name y l))  = y `elem` pl
containsProp pl  (Neg y) = containsProp pl y
containsProp pl (And x y) = (containsProp pl x) || (containsProp pl y)
containsProp pl (Or x y) = (containsProp pl x) || (containsProp pl y)
containsProp pl (Implies x y) = (containsProp pl x) || (containsProp pl y)
containsProp pl (Equiv x y) = (containsProp pl x) || (containsProp pl y)



expandAllOnce as f =  ((expandTrust as) . (expandDistrust as) . (expandWeakDistrust as)) f

expandAll as f = while (containsProp  ["Trusts","Distrusts","WDistrusts"]) (expandAllOnce as) f

expandAgents as = [Says a (expandAll as f) | Says a f <- as]
```

```
module Tableaux where

import Data.List
import AgentLogic

{- We are using Smullyans labeled tableaux. In this representation proof goals are represented by lists of labeled formula.
We adopt the names L(eft) and R(ight) for the labels (Smullyan uses T and F). We also adopt Smullyans alpha/beta notation -}

data Label a =          L a | R a
                        deriving (Show, Read, Eq, Ord)

{- data Formula a =     Prop a |
                        Neg (Formula a) |
                        And (Formula a) (Formula a)|
                        Or (Formula a) (Formula a) |
                        Implies (Formula a) (Formula a) |
                        Equiv (Formula a) (Formula a)
                        deriving (Show, Read, Eq, Ord)
-}

single (L (Neg _)) = True
single (L (And _ _)) = True
single (R (Neg _)) = True
single (R (Or _ _)) = True
single (R (Implies _ _)) = True
single (_) = False

atomic (L (Prop _ )) = True
atomic (R (Prop _ )) = True
atomic (_) = False

double (L (Or _ _)) = True
double (L (Implies _ _ )) = True
double (R (And _ _)) = True
double (L (Equiv _ _)) = True
double (R (Equiv _ _)) = True
double (_) = False

alphas (L (Neg x)) = [R x]
alphas (L (And x y)) = [L x,L y]
alphas (R (Neg x)) = [L x]
alphas (R (Or x y)) = [R x,R y]
alphas (R (Implies x y)) = [L x,R y]
alphas (x) = [x]

betas (L (Or x y)) = ([L x],[L y])
betas (L (Implies x y )) = ([R x],[L y])
betas (R (And x y)) = ([R x],[R y])
betas (L (Equiv x y)) = ([L x,L y],[R x,R y])
betas (R (Equiv x y)) = ([L x,R y],[R x,L y])
betas (_) = ([],[])

{- A goal is a list of labeled formulas.
   reduce takes a goal and transforms the goals using non-branching tableaux rules -}

reduce1 l =  concat [alphas x | x <- l]

while p f x = if p x then (while p f (f x)) else x

reduce l = nub(while (any single) reduce1 l)


dual (L x) = R x
dual (R x) = L x

closes [] = False
closes (x:l) = if (dual x) `elem` l then True else closes l


{- prove takes a list of goals and transforms the list of goals by first applying reduce, checking for brach closure
(and discarding closed branches)  and then applying branching rules to generate a new list of goals. -}

prove [] = []
prove (x:l) =
  let y = nub(reduce x)
  in    if closes y then prove l
        else if (any double y)
             then
                let (front, beta:back) = break double y
                in  let (b1,b2) = betas beta
                    in  prove   ((front++b1++back):(front++b2++ back):l)
             else y:(prove l)

entails as c = (prove [(R c):(map L as)]) == []

{-  simple uses of proves -}

{- counter writes out the the counter example as a string -}

stringList l =
        let     stringList1 [] = ""
                stringList1 [x] = x
                stringList1 (x:xs) = x ++ "," ++ (stringList1 xs)
        in
                if l == [] then "" else "(" ++ (stringList1 l) ++ ")"




counter l =
  if (all atomic l) then
        let     counter' [] = ".\n"
                counter' ((L (Prop (Name x a))):l) = x ++ (stringList a) ++ " is True\n" ++ counter' l
                counter' ((R (Prop(Name  x a))):l) = x ++ (stringList a) ++ " is False\n" ++ counter' l
        in counter' l
  else  "Not fully reduced\n"
```

```
{- pp prints the counter examples from a failed proof -}

pp l = do
         mapM_ putStrLn (map counter l)

ppp l = do
        putStrLn (if l == [] then "NO COUNTER EXAMPLES\n" else "COUNTER EXAMPLES:\n" ++ concat (map counter l))

{- functions using the parser interface -}

{- the parser uses grammar


Formula :        Formula '<=>' Formula
                | Formula '=>' Formula
                | Formula '+' Formula
                | Formula '&' Formula
                | '(' Formula ')'
                | '~' Formula
                | prop [ prop* ]
with priority ordering (least to greatest) <=>, =>, +, &, ~

E.G. A + ~A & B

is A or ((not A) and B)

-}

{- prover takes a string and returns True if the formula is a tautology and false otherwise -}

prover f = prove [ [ R (parseFormula f) ]  ]  ==  [ ]

{- consistent takes a list of strings (formula) and returns true if the list is consistent and false otherwise -}

consistent l = prove [(map (L . parseFormula) l)] /= []

{- refutation lists the counter examples of  a formula -}

refutation f =
        let l = prove [ [ R (parseFormula f) ]  ]
        in
          do
          putStrLn (if l == [] then "NO COUNTER EXAMPLES:\n"
                     else "COUNTER EXAMPLES:\n" ++ concat (map counter l ))
```

```
module Beliefs where

import PPrint
import Data.List
import Sets
import AgentLogic
import Tableaux
import Relations
import SyntacticRelations
import TrustExpansion
import Coalitions


diag [] = []
diag (x:xs) = (x,x):(diag xs)



{- variant of successor -}

equivpart ord =
        let
            eq = extractOEq ord
        in
                (closeRST eq)

{- added in diag to equ -}

extractBaseOrdSimple ass synOrd =
        let fld = [ a | Says a _ <- ass]
            ord = (successor (concat(map restructure synOrd)))
            equ = (equivpart (concat(map restructure synOrd)))
                in
                Ord fld (nub (equ++(diag fld)))  ord

extractBaseOrdTransitive ass synOrd =
        let fld = [ a | Says a _ <- ass]
            ord = (successor (concat(map restructure synOrd)))
            equ = (equivpart (concat(map restructure synOrd)))
                in
                Ord fld (nub (equ++(diag fld)))  (closeTrans ord)


extractCoaltionOrd cmp ass =
        let c = successor(coalitions cmp equSets ass)
         in
           Ord (field c) [] c



data Assertions a = AgentAssert (a, Formula) | SourceAssert ([a], Formula)
                    deriving (Show, Read, Eq, Ord)



data Conclusion a = AgentConclusion (OrderedSet a) [[a]] [Assertions a] [Assertions a] [(Formula, Bool, Bool)]  |
                    SourceConclusion (OrderedSet [a]) [[a]] [Assertions a] [Assertions a] [(Formula, Bool, Bool)]
                    deriving (Show, Read, Eq, Ord)

consistentWithHyp fs f = (f, consistentFormulaSet (f:fs), consistentFormulaSet ((Neg f):fs))


{- mod to print out maximal sets rather than ends -}




theories as ms = [ bigCup[ [f | Says a1 f <- as, a1 == a]  | a <- s] | s <-  ms ]


proofSets ths f = [  m  | m <- ths,   not(consistentFormulaSet ((Neg f):m)) && (consistentFormulaSet (f:m)) ]


{- coding in table
u-code == code for extent of undeterminedness
t-code == code for extent of truthset

(f, t-code, u-code)

if tp union tn == ts  then ths --> t-code == t else t-code = false

if either tp or tn obtains a limit (either [] or ths) then the u-code is true
otherwis eit is false.

-}


classify ths f =
        let tp = proofSets ths f
            tn = proofSets ths (Neg f)
        in
            if       (ths /= []) && (tp `equalSets` ths) && (tn == []) then (f, True, True)
            else if  (ths /= []) && (tp == []) && (tn `equalSets` ths) then  (Neg f, True, True)
            else if  (tp == []) && (tn == []) then (f, False, True)
            else if  (ths /= []) && (tp `union` tn) `equalSets` ths then (f, True, False)
            else (f, False, False)


hypClass as ms hyp =
        let ths = theories as ms
        in   [ classify ths f | f <- hyp]
```

```
beliefs s =
  let (ord, as, hy)  = parseSystem s
   in
    let
       asexp  = expandAgents as
        in
           case ord of
                (Simple ordx) -> (let
                                       orders = extractBaseOrdSimple asexp ordx
                                in
                                       let maxSets = (maximalSets asexp orders)
                                        in
                                          let agentTheory = [ AgentAssert(a, mkAnd[f | Says a1 f <- as, a1 == a])
                                                              | a <- (bigCap maxSets), [f | Says a1 f <- as, a1 == a] /= [] ]
                                               in
                                                       AgentConclusion orders
                                                                       maxSets
                                                                       [ AgentAssert(a,f) | Says a f <- asexp]
                                                                       agentTheory
                                                                       (hypClass as maxSets hy) )
                (Transitive ordx) -> (let
                                       orders = extractBaseOrdTransitive asexp ordx
                                in
                                       let maxSets = (maximalSets asexp orders)
                                        in
                                          let agentTheory = [ AgentAssert(a, mkAnd[f | Says a1 f <- as, a1 == a])
                                                              | a <- (bigCap maxSets), [f | Says a1 f <- as, a1 == a] /= [] ]
                                               in
                                                       AgentConclusion orders
                                                                       maxSets
                                                                       [ AgentAssert(a,f) | Says a f <- asexp]
                                                                       agentTheory
                                                                       (hypClass as maxSets hy) )


                (_) ->
                               (let
                                       ass    = collectSupporters (collectSupport asexp)
                                       orders =
                                               (case ord of
                                               (Majority) ->      extractCoaltionOrd majority ass
                                               (Weighted wl) ->   extractCoaltionOrd (weigh wl) ass
                                               (Upper ordl) ->    extractCoaltionOrd (upper (mkSuccOrds ordl)) ass
                                               (Lower ordl) ->    extractCoaltionOrd (lower (mkSuccOrds ordl)) ass
                                               (Combined ordl) -> extractCoaltionOrd (combined (mkSuccOrds ordl)) ass)

                                 in
                                       let sourceTheory = [ SourceAssert(c, f) | c <-  consistStar' ass (ords orders) (ends orders), (c1,f) <- ass, c == c1]
                                               in
                                                       SourceConclusion  orders
                                                                         (ends orders)
                                                                         [SourceAssert t | t <-  ass]
                                                                         sourceTheory
                                                                         [   consistentWithHyp  [f | SourceAssert (a, f) <- sourceTheory] h | h <- hy ] )
```

```
module Pretty where

import PPrint
import Data.List
import Sets
import AgentLogic
import Tableaux
import Relations
import SyntacticRelations
import TrustExpansion
import Coalitions
import Beliefs

tupledOpt [] = empty
{- tupledOpt l = tupled l -}

tupledOpt l = lparen <> (foldrOpt (<>) empty (intersperse (comma <> space) l)) <> rparen

mkDocOfName (Name a l) = (text a) <> (tupledOpt (map text l))

mkDocOfNames l = encloseSep lbrace rbrace comma (map mkDocOfName l)

mkDocOfSets ls = foldrOpt (<$>) empty (map mkDocOfNames ls)

mkDocOfNameSets l = encloseSep (text "Coalitions ")  empty comma  (map mkDocOfNames l)

priority (Prop _)       = 5
priority (Neg _)        = 5
priority (And _ _)      = 4
priority (Or _ _)       = 3
priority (Implies _ _)  = 2
priority (Equiv _ _)    = 1


parenOpt x f =  if (priority f) < (priority x) then  softline <> (parens (mkDocFormula f))
                 else (mkDocFormula f)


{-
mkDocFormula (Prop n)                   = mkDocOfName n
mkDocFormula (a @(Neg f))               = (char '~') <> (parenOpt a f)
mkDocFormula (a @(And f1 f2))           = (parenOpt a f1) <> (text " & ") <>  (parenOpt a f2)
mkDocFormula (a @(Or f1 f2))            = (parenOpt a f1) <> (text " + ") <>  (parenOpt a f2)
mkDocFormula (a @(Implies f1 f2))       = (parenOpt a f1) <> (text " => ") <> (parenOpt a f2)
mkDocFormula (a @(Equiv f1 f2))         = (parenOpt a f1) <> (text " <=> ") <> (parenOpt a f2)
-}




mkDocFormula (Prop n)                   = mkDocOfName n
mkDocFormula (a @(Neg f))               = (char '~') <> (parenOpt a f)
mkDocFormula (a @(And f1 f2))           = PPrint.group(nest 2 ((parenOpt a f1) <> (text " & ") <>  (parenOpt a f2)))
mkDocFormula (a @(Or f1 f2))            = PPrint.group(nest 2 ((parenOpt a f1) <> (text " + ") <>  (parenOpt a f2)))
mkDocFormula (a @(Implies f1 f2))       = PPrint.group(nest 2 ((parenOpt a f1) <> (text " => ") <> (parenOpt a f2)))
mkDocFormula (a @(Equiv f1 f2))         = PPrint.group(nest 2 ((parenOpt a f1) <> (text " <=> ") <> (parenOpt a f2)))




topLevelparenOpt x f =  if (priority f) < (priority x) then  (parens (mkTopLevelAnd f))
                        else (mkTopLevelAnd f)

mkTopLevelAnd (a @(And f1 f2)) = (topLevelparenOpt a  f1) <> (text "   &") <> line <> (topLevelparenOpt a  f2)
mkTopLevelAnd  f = mkDocFormula f




mkDocOfAgentOrder (a,b) = (mkDocOfName a) <> (text " > ") <> (mkDocOfName b)

mkDocOfAgentOrders [] = text "Flat Set"
mkDocOfAgentOrders l = foldr (<>) empty (intersperse (comma <$> empty)  (map mkDocOfAgentOrder l))


mkDocOfSourceOrder (a,b) = (mkDocOfNames a) <> (text " > ") <> (mkDocOfNames b)

mkDocOfSourceOrders [] = text "Flat Set"
mkDocOfSourceOrders l = foldr (<>) empty (intersperse (comma <$> empty)  (map mkDocOfSourceOrder l))

mkDocAgentAssertion (n,f) = (mkDocOfName n) <> (text ": ") <> (mkDocFormula f)

mkDocAgentAssertions l = foldr (<$>) empty (map mkDocAgentAssertion l)

mkDocOfSources ns =    encloseSep lbrace rbrace (comma <> space) (map mkDocOfNames ns)

mkDocAssertion (AgentAssert (n,f)) = (mkDocOfName n) <> (text ": ") <> (align (mkTopLevelAnd f))
mkDocAssertion (SourceAssert (ns,f)) = (mkDocOfNames ns) <> (text ": ") <> (mkTopLevelAnd f)

mkDocAssertions l = foldrOpt (<$>) empty (map mkDocAssertion l)

mkDocHypothesis (f, True, True) =   (text "scep    ")  <>  (mkDocFormula f)
mkDocHypothesis (f, True, False) =  (text "cred D  ")  <>  (mkDocFormula f)
mkDocHypothesis (f, False, True) =  (text "ND      ")  <>  (mkDocFormula f)
mkDocHypothesis (f, False, False) = (text "cred N  ")  <>  (mkDocFormula f)

mkDocHyps [] = text "No Hypothesis"
mkDocHyps l = foldrOpt (<$>) empty (map mkDocHypothesis l)

mkDocResults [] = text "No Preferred Consistent Set"
mkDocResults l = foldrOpt (<$>) empty (map mkDocAssertion l)
```

```
mkDocConclusions  (AgentConclusion orders maximal assertions result hyps) =
                                         (text "Ordering") <> line <>
                                               (indent 3 (mkDocOfAgentOrders (ords orders))) <> line <>
                                         (text "Maximal Sets") <> line <>
                                               (indent 3 (mkDocOfSets maximal)) <> line <>
                                         (text "Assertions") <> line <>
                                               (indent 3 (mkDocAssertions assertions)) <> line <>
                                         (text "Core Beliefs") <> line <>
                                               (indent 3 (mkDocResults result))  <> line <>
                                         (text "Hypothesis") <> line <>
                                               (indent 3 (mkDocHyps hyps))  <>
                                         line

mkDocConclusions (SourceConclusion orders ends assertions result hyps) =
                                         (text "Ordering") <> line <>
                                               (indent 3 (mkDocOfSourceOrders (ords orders))) <> line <>
                                         (text "Least elements") <> line <>
                                               (indent 3 (mkDocOfSources ends)) <> line <>
                                         (text "Assertions") <> line <>
                                               (indent 3 (mkDocAssertions assertions)) <> line <>
                                         (text "Core Beliefs") <> line <>
                                               (indent 3 (mkDocResults result))  <> line <>
                                         (text "Hypothesis") <> line <>
                                               (indent 3 (mkDocHyps hyps))  <>
                                         line
```

```
module Inspect where

import PPrint
import Data.List
import Sets
import AgentLogic
import Tableaux
import Relations
import SyntacticRelations
import TrustExpansion
import Coalitions
import Pretty
import Beliefs



theory s = mkDocConclusions (beliefs s)

inspect = theory
```

```
% top level driver that reads from standard in and writes to standard out.

module Main where

import Data.Char
import Inspect

main = do
                th <- getContents
                putStr (show (theory th))
```

Bibliography

[1] Leila Amgoud and Claudette Cayrol. On the acceptability of arguments in preference-based argumentation. In Gregory F. Cooper and Serafín Moral, editors, *UAI*, pages 1–7. Morgan Kaufmann, 1998. ISBN 1-55860-555-X.

[2] Leila Amgoud, Claudette Cayrol, Marie-Christine Lagasquie-Schiex, and P. Livet. On bipolarity in argumentation frameworks. *Int. J. Intell. Syst.*, 23 (10):1062–1093, 2008.

[3] Martha Amram and Nalin Kulatilaka. *Real Options: Managing Strategic Investment in an Uncertain World*. Harvard Business School Press, 1998.

[4] Robert Axelrod. *The Evolution of Co-Operation*. Penguin Books, 1990.

[5] Annette Baier. Trust and antitrust. *Ethics*, 96(2):231–260, 1986.

[6] Bernard Barber. *The Logic and Limits of Trust*. Rutgers University Press, 1983. Ref'd by gambetta-1988a near key area of interest.

[7] D. Elliott Bell and J. LaPadula. Secure computer systems: Mathematical foundations. vol. i. Technical report, MITRE, 1973.

[8] D. Elliott Bell and J. LaPadula. Secure computer systems: Mathematical foundations. vol. ii. Technical report, MITRE, 1973.

[9] D. Elliott Bell and J. LaPadula. Secure computer systems: Unified exposition and multics interpretation. Technical report, MITRE, 1976.

[10] C. Berge. *Graphs and Hypergraphs*. Elsevier Science Ltd, 1985. ISBN 0720404797.

[11] Daniel Bernoulli. Exposition of a new theory on the measurement of risk. *Econometrica*, 22(1):23–36, 1954.

[12] Philippe Besnard and Sylvie Doutre. Characterization of semantics for argument systems. In Didier Dubois, Christopher A. Welty, and Mary-Anne Williams, editors, *KR*, pages 183–193. AAAI Press, 2004. ISBN 1-57735-199-1.

[13] Philippe Besnard and Anthony Hunter. *Elements of Argumentation*. MIT press, 2008.

[14] Garrett Birkhoff. Lattice theory. In *Colloquium Publications*, volume 25. Amer. Math. Soc., 3. edition, 1967.

[15] Andrzej Blikle. A comparative review of some program verification methods. In Jozef Gruska, editor, *MFCS*, volume 53 of *Lecture Notes in Computer Science*, pages 17–33. Springer, 1977.

[16] Andrzej Blikle. *MetaSoft Primer, Towards a Metalanguage for Applied Denotational Semantics*, volume 288 of *Lecture Notes in Computer Science*. Springer, 1987. ISBN 3-540-18657-3.

[17] A. Bondarenko, P.M. Dung, R.A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93:63–101, 1997.

[18] Laurence BonJour. *The Structure of Empirical Knowledge*. Harvard University Press, 1985.

[19] Danah Michele Boyd. *Taken Out of Context American Teen Sociality in Networked Publics*. PhD thesis, University of California, Berkley, 2008.

[20] D. F. C. Brewer and M. J. Nash. The Chinese wall security policy. In *Proceedings of the 1989 IEEE Symposium on Security and Privacy*, pages 206–214, 1989.

[21] Chris Brink. Boolean modules. *Journal of Algebra*, 71:291–313, 1981.

[22] Chris Brink, Katarina Britz, and Renate A. Schmidt. Peirce algebras. *Formal Asp. Comput.*, 6(3):339–358, 1994.

[23] Alejandro Bugacov, Aram Galstyan, and Kristina Lerman. Threshold behavior in a boolean network model for sat. In Hamid R. Arabnia, Rose Joshua, and Youngsong Mun, editors, *IC-AI*, pages 87–92. CSREA Press, 2003. ISBN 1-932415-12-2.

[24] Jean Camp, Cathy McGrath, and Helen Nissenbaum. Trust: A collision of paradigms. *John F. Kennedy School of Government, Harvard University, Faculty Research Working Papers Series*, 2001.

[25] Dorwin Cartwright and Frank Harary. Structural balance: A generalisation of Heider's theory. *The Psychological Review*, 63(5):277–293, 1956.

[26] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In Lluis Godo, editor, *ECSQARU*, volume 3571 of *Lecture Notes in Computer Science*, pages 378–389. Springer, 2005. ISBN 3-540-27326-3.

[27] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. *Bipolar abstract argumentation systems*, pages 65–84. Springer Publishing Company, Incorporated, 1st edition, 2009. ISBN 0387981969, 9780387981963.

[28] James Coleman. *Foundations of Social Theory*. Belknap Press, 1990.

[29] Partha Dasgupta. Trust as a commodity. In Gambetta [40].

[30] James A. Davis. Clustering and Structural Balance in Graphs. *Human Relations*, 20:181–187, 1967.

[31] Edsger W. Dijkstra. Guarded commands, nondeterminacy and formal derivation of programs. *Communications of the ACM*, 18(8):453–457, 1975.

[32] Edsger W. Dijkstra. *A Discipline of Programming*. Prentice Hall, 1976.

[33] Edsger W. Dijkstra and Carl S. Scholten. *Predicate Calculus and Program Semantics*. Springer-Verlag, 1990.

[34] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

[35] Mustafa Emirbayer. Manifesto for a relational sociology. *The American Journal of Sociology*, 103(2, Sept 1997):281–317, 1997.

[36] M. Erne. Adjunctions and Galois Connections: Origins, History and Development. In K. Denecke, M. Erné, and S. L. Wismath, editors, *Galois Connections and Applications*, volume 565 of *Mathematics and Its Applications*, chapter 1, pages 1–138. Kluwer Academic Publishers, Dordrecht, 2004.

[37] M. Erne, J. Koslowski, A. Melton, and G. E. Strecker. A primer on Galois connections. *Annals of the New York Academy of Sciences*, 704:103–125, 1993.

[38] C. J. Everett. Closure operators and galois theory in lattices. *Transactions of the American Mathematical Society*, 55(3):514–525, 1944.

[39] David F. Ferraiolo and D. Richard Kuhn. Role-based access controls. In *15th National Computer Security Conference (1992), Baltimore MD*, pages 554–563, 1992.

[40] Diego Gambetta, editor. *Trust: Making and Breaking Cooperative Relations*. Basil Blackwell Ltd., 1988.

[41] Diego Gambetta. Can we trust trust? In *Trust: Making and Breaking Cooperative Relations* Gambetta [40].

[42] Giddens. *The consequences of Modernity*. Polity Press, 1990.

[43] Joseph A. Goguen and José Meseguer. Security policies and security models. In *IEEE Symposium on Security and Privacy*, pages 11–20, 1982.

[44] Davide Grossi. On the logic of argumentation theory. In Wiebe van der Hoek, Gal A. Kaminka, Yves Lespérance, Michael Luck, and Sandip Sen, editors, *AAMAS*, pages 409–416. IFAAMAS, 2010. ISBN 978-0-9826571-1-9.

[45] Per Hage and Frank Harary. *Structural Models in Anthropology*. Cambridge University Press 1983, 1983.

[46] Frank Harary. *Graph Theory*. Addison-Wesley, 1969.

[47] Frank Harary and Robert Z. Norman. *Structural Models: An introduction to the theory of Directed Graphs*. John Wiley and Son, 1965.

[48] Russell Hardin. *Trust & Trustworthiness*. Russell Sage Foundation, 2002.

[49] Russell Hardin. *Trust*. Polity Press, 2006.

[50] W. T. Harwood, J. A. Clark, and J. L. Jacob. Boolean coherence: Does it make sense? 2010. First Workshop on Logics for Systems Analysis (Edinburgh).

[51] W. T. Harwood, J. A. Clark, and J. L. Jacob. Networks of trust and distrust: Towards logical reputation systems. 2010. Workshop on Logics for Security (Copenhagen).

[52] W. T. Harwood, J. A. Clark, and J. L. Jacob. Conceptualising internet trust. 2010. World Internet Policy Project (WIP2) Workshop (Lisbon).

[53] W. T. Harwood, J. L. Jacob J. A. Clark, and R. I. Young. *Boolean Coherence and the ACH method*. 2010. Annual Conference of the International Technology Alliance (London).

[54] C. A. R. Hoare. Programs are predicates. In *Proc. of a Discussion Meeting of the Royal Society of London on Mathematical Logic and Programming Languages*, pages 141–155, Upper Saddle River, NJ, USA, 1985. Prentice-Hall, Inc. ISBN 0-13-561465-1. URL http://dl.acm.org/citation.cfm?id=3721.3729.

[55] Wilfred Hodges. *Logic*. Penguin Books, 1977.

[56] Martin Hollis. Penny pinching and backward induction. *The Journal of Philosophy*, 88(9):473–488, 1991. ISSN 0022362X. URL http://www.jstor.org/stable/2026602.

[57] Martin Hollis. *Trust Within Reason*. Cambridge University Press, 1998.

[58] Richard C. Jeffrey. *Formal Logic: Its Scope and Limits, 3rd ed.* McGraw-Hill Inc., 1991.

[59] Audun Jøsang and Stéphane Lo Presti. Analysing the relationship between risk and trust. In *Proceedings of the 2nd International Conference on Trust Management, 2004*, 2004. http://brage.unik.no/people/josang/papers/JLoP2004-iTrust.pdf.

[60] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decis. Support Syst.*, 43(2): 618–644, 2007. ISSN 0167-9236. doi: http://dx.doi.org/10.1016/j.dss.2005.05.019.

[61] Richards J Heuer Jr. *Psychology of Intelligence Analysis*. Nova Biomedical, 1999.

[62] Frank H. Knight. *Risk, Uncertainty, and Profit*. Houghton Mifflin, 1921.

[63] Loren M. Kohnfelder. *Towards a Practical Public-key Cryptosystem*. MIT Dissertation, 1978.

[64] Andrew H. Kydd. *Trust and Mistrust in International Relations*. Princeton University Press, 2008.

[65] J. David Lewis and Andrew Weigert. Trust as a social reality. *Social Forces*, 63(4):967–985, 1985.

[66] Niklas Luhmann. *Trust and Power*. John Wiley & Sons, 1979.

[67] Niklas Luhmann. Familiarity, confidence, trust: Problems and alternatives. In Gambetta [40].

[68] Stephen Paul Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, Dept. of Computing and Mathematics, University of Stirling, 1994.

[69] Daniel J. McAllister. Affect- and cognition- based trust as foundations for interpersonal cooperation in organisations. *Academy of Management Journal*, 38:24–59, 1995.

[70] Edward J. McCluskey, Kenneth P. Parker, and John J. Shedletsky. Boolean network probabilities and network design. *IEEE Transactions on Computers*, 27(2):187–189, 1978.

[71] D. H. McKnight and N. L. Chervany. The meaning of trust. *Technical Report MISRC Working Paper Series 96-04, University of Minnesota, Management Information Systems Reseach Center, 1996.*, 1996. http://www.misrc.umn.edu/wpaper/WorkingPapers/9604.pdf.

[72] D. H. McKnight and N. L. Chervany. Trust and distrust definitions: One bite at a time. In *R. Falcone, M. Singh, and Y.-H. Tan (Eds.): Trust in Cyber-societies, LNAI 2246, pp. 27 - 54, 2001*, 2001.

[73] D. H. McKnight and N. L. Chervany. What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. *International Journal of Electronic Commerce, Winter 2001 - 2002, Vol. 6, No. 2, pp. 35 - 59*, 2002.

[74] D. Harrison McKnight, Larry L. Cummings, and Norman L. Chervany. Initial trust formation in new organizational relationships. *Academy of Management Review*, 23(3):473–490, 1998.

[75] J. W. Milnor. Games against nature. In C.H. Coombs and R.L. Davis, editors, *Decision Processes*. Wiley, 1954.

[76] Barabra A. Misztal. *Trust in Modern Societies*. Polity Press, 1996.

[77] Sanjay Modgil and Trevor J. M. Bench-Capon. Integrating object and meta-level value based argumentation. In Philippe Besnard, Sylvie Doutre, and Anthony Hunter, editors, *COMMA*, volume 172 of *Frontiers in Artificial Intelligence and Applications*, pages 240–251. IOS Press, 2008. ISBN 978-1-58603-859-5.

[78] Helen Nissenbaum. *Privacy in Context:Technology, Policy and the Integrity of Social Life*. Stanford University Press, 2010.

[79] Martin A. Norwak and Karl Sigmund. Tit for tat in heterogeneous populations. *Nature*, 355(16 January):250–253, 1992.

[80] Ryan O'Donnell. Some topics in analysis of boolean functions. In Cynthia Dwork, editor, *STOC*, pages 569–578. ACM, 2008. ISBN 978-1-60558-047-0.

[81] Kieron O'Hara. *Trust: From Socrates to Spin*. Icon Books, 2004.

[82] K. Ono. On some properties of binary relations. *Nagoya Mathematical Journal*, 12:161–170, 1957.

[83] Oystein Ore. Galois connexions. *Transactions of the American Mathematical Society*, 55(3):493–513, 1944.

[84] Oystein Ore. *Theory of Graphs, 3rd edition*. American Mathematical Society, 1969.

[85] Ewa Orlowska and Joanna Golińska Pilarek. *Dual Tableaux: Foundations, Methodology, Case Studies*. Springer, 2011.

[86] Kenneth P. Parker and Edward J. McCluskey. Probabilistic treatment of general combinational networks. *IEEE Transactions on Computers*, 24(6):668–670, 1975.

[87] Kenneth P. Parker and Edward J. McCluskey. Analysis of logic circuits with faults using input signal probabilities. *IEEE Transactions on Computers*, 24(5): 573–578, 1975.

[88] L. S. Penrose. The elementary statistics of majority voting. *Journal of the Royal Statistical Society*, 109(1):53–57, 1946.

[89] Nancy Nyquist Potter. *How Can I be trusted: A Virtue Theory of Trustworthiness*. Rowman & Littlefield, 2002.

[90] Graham Priest. *An Introduction to Non-Classical Logic: From If to Is. 2nd edition*. Cambridge University Press, 2008.

[91] Robert D. Putnam. *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, 2000.

[92] Robert D. Putnam. Bowling alone: America's declining social capital. *Journal of Democracy*, 6(1):65–78, 2001.

[93] J. Riguet. Relations binaires, fermetures, correspondances de Galois. *Bulletin de la Société Mathématique de France*, 76:114–155, 1948.

[94] Andrea Roli and Michela Milano. Solving the satisfiability problem through boolean networks. In Evelina Lamma and Paola Mello, editors, *AI*IA*, pages 72–83. Springer-Verlag, 2000. ISBN 3-540-67350-4.

[95] Jean-Jacques Rousseau. *A Discourse on Inequality*. Penguin Classics, 1982. Original Publication 1754.

[96] Peter Y. A. Ryan. Mathematical models of computer security. In Riccardo Focardi and Roberto Gorrieri, editors, *FOSAD*, volume 2171 of *Lecture Notes in Computer Science*, pages 1–62. Springer, 2000. ISBN 3-540-42896-8.

[97] Karem A. Sakallah and João Marques-Silva. Anatomy and empirical evaluation of modern sat solvers. *Bulletin of the EATCS*, 103:96–121, 2011.

[98] John G. Sanderson. *A Relational Theory of Computing*, volume 82 of *Lecture Notes in Computer Science*. Springer, 1980. ISBN 3-540-09987-5.

[99] Marco Santoro. Framing notes: An introduction to "catnets". *Sociologica*, 1, 2008.

[100] Thomas J. Schaefer. The complexity of satisfiability problems. In *Proceedings of the tenth Annual ACM Symposium on Theory of Computing*, STOC '78, pages 216–226, New York, NY, USA, 1978. ACM. doi: 10.1145/800133.804350. URL `http://doi.acm.org/10.1145/800133.804350`.

[101] Gunther Schmidt. *Relational Mathematics*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2010.

[102] R. A. Schmidt. Representations as full Peirce algebras: Extended abstract. Manuscript., November 1994. URL `http://www.cs.man.ac.uk/~schmidt/publications/AMAST95extAbstr.dvi.gz`.

[103] R. A. Schmidt. Peirce algebras and relation algebras are equipollent. Manuscript., Sept. 1993/Aug. 1994. URL `http://www.cs.man.ac.uk/~schmidt/publications/equipol.dvi.gz`.

[104] Reinhard Selten. The chain store paradox. *Theory and Decision*, 9:127–159, 1978.

[105] Raymond M. Smullyan. *First-Order Logic*. New York [Etc.]Springer-Verlag, 1968.

[106] Raymond M. Smullyan. *What is the Name of This Book?: The Riddle of Dracula and Other Logical Puzzles*. Penguin Books Ltd., 1990.

[107] M. Smyth. Power domains and predicate transformers: A topological view. In Josep Diaz, editor, *Automata, Languages and Programming*, volume 154 of *Lecture Notes in Computer Science*, pages 662–675. Springer Berlin / Heidelberg, 1983. ISBN 978-3-540-12317-0. URL `http://dx.doi.org/10.1007/BFb0036946`. 10.1007/BFb0036946.

[108] Wolfgang Sofsky. *Privacy: A Manifessto*. Princeton University Press, 2008.

[109] Daniel J Solove. *Understanding Privacy*. Harvard University Press, 2009.

[110] J. Michael Spivey. *Z Notation: A Reference Manual (2nd ed.)*. Prentice Hall International Series in Computer Science. Prentice Hall, 1992. ISBN 978-0-13-978529-0.

[111] Goran Sundholm. Systems of deduction. In Dov Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic. Volume 1: Elements of Classical Logic*, pages 133–188. Dordrecht et al. Reidel, 1993.

[112] M. E. Szabo, editor. *The collected papers of Gerhard Gentzen*. Elsevier Science Publishing, 1969.

[113] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, 1994.

[114] Harrison C. White. Notes on the constituents of social structure: Soc. rel. 10, spring 1965. *Sociologica*, 1, 2008.

Index