Deception Detection using Human Reasoning

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

by

Deqing Li

Thayer School of Engineering

Dartmouth College

Hanover, New Hampshire

May, 2013

Examining Committee:

Chairman_____
Eugene Santos, Jr.

Member_____
George Cybenko

Member_____
Mark Borsuk

Member_____
Hien Nguyen

_____
F. Jon Kull
Dean of Graduate Studies

Abstract

Scientific methods that detect deception have been studied since 1895. Among them, computational methods have gained popularity in recent decades thanks to the development of artificial intelligence, (AI). Detecting deception by categorizing verbal/non-verbal cues using machine learning techniques has been the main stream approach in the field. We investigate deception detection methods that work on communication content in a written format. In this work, we propose that verbal/non-verbal cues are simply artifacts created during the implementation of deception, and we instead study the cognitive process behind the formation of deception. We detect deceptive communication by modeling the cognitive process in deception, comparing the semantic structure of deceptive communication with that of honest communication, and identifying the patterns for deceptive reasoning. Our method differs from existing works by targeting at malicious intent instead of wrong information, by deriving observations directly from the intent to deceive, and by taking individual difference into consideration. As a result we are able to distinguish unintentional misinformation from intentional deception, an approach which no existing research has yet addressed.

In representing the reasoning process of human communication we use Bayesian Networks. The contributions of our work lie with (i) its development of an alternative method of deception detection and improvement of detection performance by using the cognitive process in human argumentation, (ii) its exploration of the deep cognitive process in human argumentation through linguistic information, (iii) its ability to explain the way that a deceptive communication is formed and detected, (iv) its intuitive

representation of deceptive reasoning, which facilitates the corresponding explanation of the verbal cues of deception, and (v) its analysis of the impact of different types of deception datasets on the detection performance. We propose to compare our approach with verbal cues in terms of accuracy and reliability. Our ultimate goal is to obtain a better understanding of how humans reason, the decisions they make when they decide to deceive.

Acknowledgements

I joined the lab because of my great interest in Artificial Intelligence. I am fascinated by how human brain functions, how people think and behave and how machines can simulate human intelligence.

This thesis is a summary of the research that I have been devoted to for almost six years. This topic, deception detection, was started as a follow-up of a funded research. At first deception was just an interesting topic that attracted both researchers and laymen's attentions. As I dug into this topic deeper and longer, I figured out that it is a tremendously huge area with very long history of research. However, most of the methods to tackle the problem are disappointingly unsuccessful in their results. This motivated me to think the problem from a new direction, by looking at the intent to deceive instead of the usual approaches of catching leaked cues.

Deception is one of the most complicated and amazing brain activities that humans can perform, the understanding of which requires me to study all aspects of human intent. Fortunately, the involvement of different research projects in our lab equipped me with abundant knowledge about the research on human intent and human behavior, and with the capability to discover research problems and challenges from a topic, derive big pictures of a problem, obtain inspirations from other intellectuals, and verify ideas through implementation.

I would like to express my most heartfelt gratefulness to my supervisor Prof. Eugene Santos, Jr., who has patiently instructed me, inspired me and helped me. He always quickly identifies the crux of the problems in my research and directs me to the right area. His influence to me is not only in the research field but also on my personality. Through the training, I was able to overcome the weakness in my personality.

I would also like to thank my committee member Prof. George Cybenko, who gave me instructions and lectures on related technologies, Prof. Mark Borsuk, who spent his valuable time to provide me with constructive comments and advice on my thesis, and Prof. Nguyen Hien, who I cooperated with for several research projects and from whom I learned how to tackle a research problem. My lab mates, Fei Yu, Qi Gu, Jacob Russell, et al. also offered me tremendous help when I reached bottlenecks in the research and in life, for which I am very grateful. I also want to thank the former students and lab members who contributed to the earliest work of the research (Qunhua Zhao, Gregory Johnson, Jr., Hien Nguyen, Paul Thompson, Xiuqing Yuan). They inspired me and led me into this area of research. When preparing this thesis, I was fortunate to have Ms. Louise A. Cullen to help me carefully proofread it. Thank you, Louise. Last, but not least, my mother, who flew from China four times to accompany me during my study and gave me great comfort and encouragement. I could not have done this without the help from the people I met in Dartmouth College. For my mother and my beloved friends, I dedicate this work.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Interpersonal Detection Theory	IDT
Artificial Intelligence	AI
Part of Speech	POS
Linguistic Inquiry and Word Count	LIWC
Adversary Intent Inference	AII
Bayesian Networks	BNs
most probable explanation	MPE
standard	std.
Support Vector Machine	SVM
Natural Language Processing	NLP
Amazon Mechanical Turk	AMT
conditional probability tables	CPTs
random variables	r.v.s
maximum likelihood estimates	MLE
area under ROC curve	AUC

Chapter 1 Introduction

1.1 Overview

Human beings develop the skill of deception in early childhood (Ford, 1996), and improve the skill of deception through practice as they age. By recognizing the effect of his words or actions on the receiver's beliefs, an experienced deceiver can manipulate the behavior of another person. Deceiving, in most cases, is a malicious act that by altering receivers' knowledge through false information can bring long-term and irreparable harm to receivers. However, deception detection is extremely difficult. Human deception detection skills have been found to only be slightly better than chance – more precisely, 45% to 65%, thus implying that human detection skills are typically no better than random guessing (Millar and Millar, 1997). When faced with lies on random topics, professional detectors such as police and customs inspectors perform no better than non-professional detectors (Bond and DePaulo, 2008). Past research that reported the existence of competent human detectors was found to be flawed by statistical error (Bond and Depaulo, 2008). Although detection rate can be slightly boosted by being more aware of deception, detection skills cannot be significantly improved through training because general heuristics cannot be applied to specific individuals (Ford, 1996; Johnson et. al, 2001). Johnson et al. (2001) noticed that people learn knowledge about how to apply detection heuristics from past experience only if a particular form of deception is frequent. However, deception detection is a low base-rate task as deception occurs infrequently, especially in domains where face-to-face interactions and feedback are available to

facilitate truth-telling (Johnson et al., 2001). Therefore, people's experience in detecting deception is fraught with failure.

In modern times, deception detection becomes an even more critical capability since modern communication technologies, such as online social networks, enable people to communicate in a wider scope at a faster pace. Information exchange may be performed using anonymous identities and without geographical as well as (sometimes) language indicators. In a society where different sectors are closely connected, deceptive information can spread quite rapidly with attendant negative consequences. Furthermore, communication through the internet such as via emails, blogs and twitters also eliminates leakage of nonverbal cues while delaying the need for immediate unprepared responses – which detectors heavily rely on. Additionally, communication over long distance with little social presence may increase the risk of deception by encouraging norm-breaking behaviors (George and Carlson, 1999). Evidence shows that people's average detection rate in computer-mediated communications is worse than a random guessing (Zhou and Zhang, 2012). This makes the development of a technology that can accurately and reliably detect deception in our modern context an urgent and important task.

To have a machine accomplish a cognitive function that even humans cannot do well is very challenging, but fortunately, machines are able to overcome certain obstacles that deter humans from uncovering truth. First of all, humans tend to believe in whatever they

want to believe rather than the facts. According to Ford (1996), when new information is received by the human brain, it is checked against old information, and registered if it is inconsistent with the old information. However, if the new information is incompatible with desires, it will not reach our consciousness, nor activate alarms that may alert the receiver to be aware of deceptions. On the other hand, machines do not conceal information. They store information completely, and detect/calculate inconsistency precisely.

Interpersonal Detection Theory (IDT) (Buller and Burgoon, 1994) illustrates that a human's cognition and behavior during a deceptive interchange varies according to the role of and/or the listener's relationship with the speaker – supposing that the person who gives information is the "speaker" and the person who receives information is the "listener". People are highly inclined to trust the communication of their relational partners, which is defined as *truth-bias* by McCornack and Levine (1990). *Social Facilitation Theory* (Zajonc, 1969) suggests individuals perform differently if there are others involved compared to how they perform if alone. Consequently, listeners perform poorly in deception detection because they are involved in communication with deceivers (George and Marett, 2004). In contrast to humans, machines are not influenced by the social context and can judge each individual interaction impersonally. The drawback of machines is their lack of ability to handle the subtleness and ambiguity in linguistic content. However, the rapid development in AI has started to equip machines with human-like reading and comprehension capabilities so that they can process the

semantics of natural language for uses such as morphological analysis that determines words, nonword tokens, and parts of speech in a sentence, shallow syntactic parsing that identifies some phrasal constituents without indicating their internal structures and functions in a sentence, and lexical semantic analysis interprets the meaning of words without resolving the entire sentence's meaning, to name a few (Zhou and Zhang, 2012). Historically, the first machine that assisted deception detection can be traced back to 1895, when a device was invented to measure the change in blood pressure for police cases (Inbau, 1948). To detect physiological responses to deception, machines such as polygraphs and functional MRIs are now used by professional polygraphers. Although the accuracy can be as high as 100% when multiple machines agree, the cost was shown to be extremely high. Besides, invasive methods that measure physiological signals aren't so easy to apply in group conversations (Hung, 2012).

By observing the detection methods of successful human detectors, researchers posited that clues in communication channels such as words, cadence, volume and pitch of the voice, facial expressions, movements or posturing of the trunk and limbs, and observable physiological reactions to emotion are effective indicators of deceivers. Professional poker players, who are among the most competent human detectors (Ford, 1996), are able to detect deception by observing opponents' gestures and facial expressions. In order to find clues as to the cards their opponents are holding and the actions they may take, the poker players rely on catching such clues. Use of these behavioral cues or "non-verbal cues" has gained popularity among researchers for the past several decades. However, in

long-distance communication non-verbal cues are more and more difficult to capture due to the less frequent face-to-face communications.

It was only recently that people have started to look at the verbal cues for detecting deception. Verbal cues refer to the subjective and objective features related to the wording and phrasing patterns in the content of communication - quantity, nonimmediacy, diversity, specificity, language complexity, cognitive complexity, informality, expressivity, affect, and uncertainty, to name a few. Psychologists and linguistics have found that deceivers tend to use words and phrases with certain cognitive features more often than truth tellers. These observations can be tied to their psychological states. DePaulo etc. (2003) provided a comprehensive review of both verbal and non-verbal cues that have been used by researchers. Despite all the successes with verbal-cues, the reliability of verbal-cues for deception detection is still questionable. For example, it is not difficult for a deceiver to change the wording once the deceiver knows the triggering words or if the deceiver has time to prepare for a response especially if the deceiver is communicating by asynchronous communications. Asynchronous communication refers to written communication that can be planned such as reports and emails. Synchronous communication refers to more instant communication such as instant messaging and audio conferencing. Also, wording styles are different from person to person and are mediated by culture. Lastly the framing of the question can also influence the framing of the response (e.g., the answer to the question “Did you go to class today?” always starts with “Yes, I” or “No, I”). Thus, detection using verbal cues is in general limited to

informal and unplanned communications, and generally results in detecting typically daily lies. In this thesis, we focus on formal and asynchronous communications since they are where conventional detection technologies such as verbal cues are not applicable and where deceivers could cause catastrophic damage without being questioned immediately such as financial losses caused through cognitive hacking (Cybenko et. al, 2002), behavior manipulation targeted by propaganda (Miller, 2004), and security risk posed by malicious insiders (Santos et al., 2012).

1.2 Problem Definition

Over the centuries, in order to detect deception people have been completely dependent on the clues that deceivers "leak" during deceptive interchanges. It seems that detecting deception from the lie itself is a "mission impossible." Researchers have a deep understanding of why and how humans deceive, but the ability to deceive does not seem to transfer to the ability to detect deception, and no person or method has shown to be dominantly more successful than others. In this dissertation, we propose a method that detects deception by ascertaining and then detecting the reasoning of deceivers. Consider the strategy that the thieves used to rob the vault in the movie *Ocean's Eleven* (2001). In the movie, the thieves silently broke into the vault, but at the same time revealed to the owner that he was being robbed in order to negotiate with him. They threatened to blow up half of the money in the vault if the owner did not let them leave with the other half. Note that the negotiation is very risky as a rational owner is unlikely to accept the negotiation in order to prevent future crimes. As such, it appears to be unreasonable for

the thieves to take only half of the money with such a risky negotiation. There must be a better explanation for such negotiations assuming the thieves are rational. In the movie, the owner called the police and a confrontation in the vault resulted in an explosion that "destroyed" all the money. The result seems unfavorable to both the owner and the thieves. However, given that the thieves had expected the owner to call the police, and that the only people who could freely move in and out of the vault are police officers, it is reasonable to assume that the thieves disguised themselves as the police in order to walk out with the money. Reasoning in this way, the negotiation, the confrontation in the vault and the explosion of the money can all be explained. As we can see from this example, investigating the reasoning of deceivers by connecting all the observations and finding the best explanation from deceivers' perspective can help identify deception and recover truth. Although the deceivers try to hide their true intent they cannot avoid showing some unexpected/inconsistent behaviors such as the aforementioned negotiations because of the deceivers' goal of leaving the vault safely. Furthermore, the whole of their behaviors can be better explained by a malicious hypothesis (which is to steal all the money as well as leave safely) than with the deceptive hypothesis (which is to steal only half the money and take a risky strategy).

1.3 Scope of Deception

Many definitions of deception arise from numerous disciplines and situations studied (Whaley, 1982; Burgoon and Buller, 1994; Baron 1988; Barnes 2007; Mahon 2008). Whaley (1982) defines deception as information designed to "manipulate the behavior of

others by inducing them to accept a false or distorted presentation of their environment-physical, social, or political", and Burgoon and Buller (1994) defines deception as a "deliberate act perpetrated by a sender to engender in a receiver's beliefs contrary to what the sender believes is true to put the receiver at a disadvantage". The definition has been argued by philosophers over the centuries. The main argument focuses on the intentionality of deception and the making of a false statement. Although a number of philosophers hold that deceiving may be unintentional (Demos 1960; Chisholm and Feehan 1977; Adler 1997; Gert 1998; Fuller 2008), many others have argued that it is not possible to deceive unintentionally (Linsky 1963; Van Horne 1981; Barnes 2007). The objections on the making of a false statement claim that any statement made with an intention to deceive is a lie (Bok 2001; Barnes 1994; Davidson 1987). Other arguments are centered on the success of deception (Ryle 1949), the speaker's belief on the false information (Linsky 1963; Fuller 2008; Schmitt 1988; Barnes 2007) and whether the making of a statement is necessary for deception (Vrij 2008; O'Neill 2003; Ekman 2009; Scott 1994). Due to the large disagreement on the definition of deception, the purpose, in this section, is to define the scope of deception that our research is focused on. Our concept of deception is defined in particular as follows (Burgoon and Buller, 1994):

In a deceptive communication,

- The information is false from the speaker's point of view.
- The act is intentional.
- The purpose is to take advantage of the listener.

As researchers suggest (Pinto, 2001; Levi, 1996), attitudes are beliefs with varying degrees of confidence. Therefore, it is more reasonable to represent a speaker's point of view as belief with uncertainty. In this sense, the "information" in the first bullet refers not only to the speaker's polarity of belief, but also to his degree of belief. It follows that any significant deviation from one's degree of belief can be deception. For example, hiding can be a deviation from certain belief (knowing something for sure) to uncertain belief (being indifferent to options) and fabrication can be a deviation from uncertain belief to certain belief. The second bullet constrains deception to intentional behavior. Intentional behavior refers to "action that the agent does for a reason" (Goldman, 1970), and it is impossible to perform an intentional action without some appropriate reason (Davidson, 1980). Thus, assuming an agent carries out an action for the reason *A*, the components of his intentional behavior are: the desire of *A* (motivational aspect), the belief that he will do *A* (settling aspect), and a plan *B* to realize the desire (representational aspect) (Mele, 1992). In the context of reasoning, reasoning based on arguments is not an intentional behavior because the belief in the arguments is not a desire, but a perception of the world. On the other hand, reasoning according to conclusion is an intentional behavior because it is driven by a desire to convince the listener of the validity of a specific conclusion, a belief that the speaker will carry out the conversation, and a plan to do it by deriving convincing arguments. Intention was also discussed with the concept of "direction of fit" (Humberstone, 1992). Direction of fit distinguishes two attitudes relating propositions to the world (Velleman, 1992): beliefs, that aim at the truth and so aim to fit the world in a *mind-to-world* direction, and desires, that normally express a yet to be realized state of affairs and so have a *world-to-mind*

direction of fit. An intentional act, from Velleman's view (1992), simultaneously contains both the mind-to-world and the world-to-mind directions of fit. This concept again confirms that inferring conclusion from arguments is not an intentional behavior because arguments are a speaker's beliefs about the world. Reasoning from arguments conceives the world in a mind-to-world direction. World-to-mind fit occurs when a speaker claims that his conclusion is true regardless of the observations. If his arguments reasonably support the claimed conclusion, a mind-to-world fit is also presented. Derived from a desired conclusion, deception belongs to an intentional behavior. The third bullet specifies that deception is a malicious behavior. By defining deception in this way, we exclude daily lies such as white lies, mental disease, self-deception and children's blatant deceptions. Deceptions that fall in this definition are more interesting for research purposes because they are more damaging and more common in professional arenas.

Deception can be categorized based on the cognitive load that is put on the deceiver. In first-order deception, the deceiver is aware of the listener's beliefs. In second-order deception, the deceiver considers the listener's evaluation of the speaker's own belief. Third-order deception considers the listener's evaluation of how the speaker would evaluate his belief, and so on. In our thesis, we restrict ourselves to first order deception since first order deception is most common. Higher order deception is the focus of interpersonal deception, the detection of which requires knowledge of the listener's intent in addition to the speaker's communication content. Moreover, we assume that the speaker's communication content does not change according to feedback. Although IDT

argues that speakers react to the listeners' suspicion display, the theory provoked a number of disagreements. Research that tried to validate this theory through empirical studies (Vasilyeva and Frank, 2011) reached the contradictory conclusion that deceivers do not always monitor their interlocutors' behavior. Instead, they found that most deceivers ignore the feedback from the other party and stick with their preplanned strategies as the communication moves on. Lastly, we require that a sufficient level of reasoning is involved and presented in the communication content, as opposed to communication with only "yes/no" answers or communication that simply expresses emotions.

Deception has also been categorized based on the technique used for the deception, the purpose of the deception and the tactics of the deceiver. The most commonly accepted taxonomy of deception belongs to Bell and Whaley (1991). In this research, we do not intend to develop a new taxonomy of deception, but we plan to propose a model that has the potential to be general enough to categorize deception based on the existing taxonomy.

Some unintentional deviations from one's original beliefs (referred to as misinformation) can be easily confused with deception, such as opinion change, innovation, and misinformation. They should not be regarded as deception because they do not possess all three elements in the definition of deception. In this dissertation we plan to distinguish

the non-deceptive misinformation from deception by hypothesizing that deceivers' reasoning patterns are unique to deception but not to other unintentional misinformation.

1.4 A Review of Computational Methods of Deception Detection

Computational methods of deception detection are mainly focused on three types of approaches. In one method, people search for verbal and non-verbal cues in the content of deceptive interchange. In the second, people consider deception types and taxonomies, and use different counter strategies to detect and reverse deception. In the third, social networks are used to evaluate the trust and reputations of agents. We will discuss each type of the approach in this section.

To find mechanisms that help detect deception, researchers looked for behavioral cues from animal behavior (Wile, 1942), child behavior (Sodian, 1991), military behavior (Cruickshank, 1979), athletes behavior (Brault, et al., 2012) and internet behavior (Grazioli and Jarvenpaa, 2000). Some researchers attempted to study both behavioral and verbal content, which offered them more evidence for identifying potential conflicts or inconsistencies. Heuristics are applied to measure the cues in the verbal and behavioral channels both qualitatively and computationally. Qualitative evaluations include cues from verbal content such as the action of depersonalizing one's answer by offering his belief on the subject instead of answering directly, cues from the manner of communication such as reactions that are out of proportion to the question, and cues from

deceivers' psychological state, such as the tendency to suspect others. Computational measurements, on the other hand, carefully encode the cues into numerical values and use statistical tools to measure significance. DePaulo et al. (2003) summarized most of the existing works on deceivers' computational cues and placed them into five categories: (i) liars are less forthcoming than truth tellers, (ii) liars tell less compelling tales than truth tellers, (iii) liars are less positive and pleasant than truth tellers, (iv) liars are more tense than truth tellers, and (v) liars include fewer ordinary imperfections and unusual contents than truths tellers. Specific and quantifiable cues under each category are evaluated according to their significance to detection. The combination of significant cues can achieve an average detection rate of 70%. However, a large number of non-verbal cues are micro-expressions that last less than one-fifth of a second. They can be easily overlooked by human detectors. To elicit more emotional responses from deceivers, rules were proposed as to how to question deceivers (Lieberman, 1999), but the effect is yet to be tested. With people relying more frequently on communications through media, detection based on verbal content becomes more critical. Verbal cues are intuitive indicators of deception because the heavy cognitive load and the psychological reactance of deceivers make them use different words and compose different sentences during a deceptive interchange. Verbal cues are also easy to retrieve. Most of the modern communications including audio, video and written ones are recorded in one form or another. Verbal cues are simply retrieved from the appearance of words and attributes of words such as sentiment of words. Then classification tools from machine learning and statistics are used to classify deceptive stories based on the attributes. Many studies have proved the accuracy of verbal cues and found explanations of the cues in psychology and

cognitive science. Mihalcea et al. (2009), as one of the pioneers in the study of verbal cues, proposed ten classes of words that have the most discriminative power to classify texts to deception/truth. They observe that among all discriminative classes, the classes of human-related word show detachment from the self in deceptive stories and close connection to the self in honest stories; words related to certainty are more dominant in deceptive stories; and the belief-oriented vocabulary is more indicative of truth. The detection performance of verbal cueing varies between 65% and 75% according to the topic of a dataset. However, verbal cues in stories of one topic do not apply to stories of another topic.

Despite the recognized success of verbal and non-verbal cues, we find these approaches to be limiting and potentially problematic. First of all, they hold the wrong assumption that heuristics derived from the general population can be applied to individuals. In cueing methods, individuals are always compared with the general population. Since words rarely used by the majority of truth tellers are identified as "discriminative words" for deception, whoever happens to frequently use these "discriminative words" are always classified as deceivers. This is incorrect because according to the definitions of deception, deception is a wrong belief from the deceiver's point of view. It is not related to what the general population believes or what is "rational" according to the public standard but only depends on the speaker himself. Pinto (2001) says: "when people cannot be expected to realize that their premises are inconsistent, their guilt is not so clear." It means that if a person truly believes in what he argues no matter whether it's

true or not in reality, he is not guilty of deception. For example, a person who does not believe in the existence of god but claimed himself to be a Christian was not deceiving if he truly believes that Christians do not assume the existence of god. Thus, using general rules to identify deceiver should be avoided, and it should only be the last choice when personal information about the speaker is unavailable. On the contrary, an effective detection method should measure the deviation from the self in addition to the deviation from the general population. Secondly, the intent of the speaker cannot be revealed from verbal or non-verbal cues. Although deceivers are found to be more cognitively challenged, guilty, anxious or insecure than a person who is telling the truth (DePaulo et al., 1985), not everyone physiologically reacts the same to anxiety-provoking situations (Ford, 1996; Hung, 2012). The intent of frequent blinking can be nervousness, the intent of pauses in sentences can be heavy mental activities, and the intent of increased pitch of voice can be emotional change. These physiological actions may be observed more often in deceptive communication, but they are neither sufficient nor necessary indicators of a malicious intent of deception. If we refer back to the three elements of the definition of deception (Burgoon and Buller, 1994) (the information is false from the speaker's point of view, the act is intentional, and the purpose is to take advantage of the listener), it is not difficult to conclude that deception is not the act of showing abnormal behavior itself, but a malicious intent that may result in unexpected behavior. What follows this definition is that unintentional deviations from the self such as misunderstanding, wrong assumptions and changed opinions are excluded from the definition of deceptive communication. However, none of the existing methods or deception datasets attempt to discriminate unintentional errors from intentional deception. Another concern occurs to

us following the concept that deception is driven by intent. Verbal cues are not the direct products of deceptive intent but the artifacts, and some non-verbal cues are merely the by-products of deception. As an artifact, the generation of verbal cues has been mediated by the mental state of the speaker and the communication environment, such as the wording style of the speaker and the framing of the question. It is hard to eliminate the "seasoning" from these processes. As a result, cueing methods only apply to a limited number of situations, and the detection heuristics are obtained from trial and error as opposed to derived by fundamental theories. To reveal the true intent of a deceiver from the communication content, we need to move closer to his mental process than verbal cues seem to allow.

Methods beyond word-level detection include part of speech (POS) tags, *Linguistic Inquiry and Word Count* (LIWC) (Pennebaker et al., 2007), and rhetorical structures (Rubin and Vashchilko, 2012). In Ott et al. (2011), the authors found that in addition to wording, POS also reveals the writing style of an individual, and performed genre identification through the frequency distribution of POS tags in a text. LIWC is a computerized text analysis program that outputs the percentage of words in a given text that fall into one or multiple linguistic (e.g., the rate of misspelling), psychological (e.g., anger, achievement), personal (e.g., leisure, money) and spoken (e.g., filler and agreement words) categories. LIWC is widely used in social science to analyze the cognitive features of texts, and is used in (Newman et al., 2003; Ott et al., 2011; Hancock et al., 2007) for detection deception. Deceptive texts, from Rubin and Vashchilko's view

(Rubin and Vashchilko, 2012), can be identified by their rhetorical structures, that is, relationship between sentences, thus, they proposed a model that incorporates sentence structures and text coherence analysis in the interpretation of communication. The attributes of text-level and cognition-level may be more reliable than word-level features under variable environrmments, but word-level detections still remain to be the most accurate method. Since communication starts with the formation of arguments, it is reasonable to assume that deception is rooted in the formation of deceptive arguments. Unfortunately, in reviewing the literature, we have not seen a cueing method that is closely connected with the intent or the formation of deception no matter whether they classify based on words, sentence structures or cognitive features. To develop an accurate and reliable model of deception detection, we claim that researchers need to refer to the knowledge of deceivers and the process of deceiving in order to identify the essential uniqueness of deceivers.

To the best of our knowledge, although the problem of neglecting personal difference has been pointed out in recent research there is not yet a good solution to it. The problems with respect to the intent of deception are slightly touched upon by another school of research that identifies deception by the tactics of deceivers, although not yet deeply studied and discussed. The most compelling work in this school is the taxonomy of deception by Whaley and Bowyer (Bell and Whaley, 1991; Bowyer, 1982), who proposed to categorize deception based on the goals of deceivers. According to (Bell and Whaley, 1991), deception can be categorized into simulative deception and dissimulative

deception. The aim of simulative deception is to create false beliefs. It is thus further split into *Mimicking*, *Inventing*, and *Decoying*. Dissimulative deception focuses on hiding the truth, and is divided into *Masking*, *Repackaging*, and *Dazzling*. This taxonomy of deception is very important. It provides a theoretical basis to the research that studies the classification/identification of deception tactics. For example, Johnson et. al (2001) examined the way that auditors detect malicious manipulations of financial information by management who purposefully make the company appear more profitable than it actually is. Inspired by the detection strategy of auditors, Johnson et. al suggest the use of four processes to detect deceptions. Firstly detectors compare expectations with observed values. The magnitude of the discrepancy between them determines whether to activate further checks. If further checks are needed, detectors attempt to generate deception hypotheses about suspected manipulations in the environment. The hypotheses are generated according to the domain knowledge of financial fraud and the taxonomy of deception. Then, in order to evaluate the hypotheses, the representation of the observed environment is edited so that it is consistent with the hypothesized manipulations. Finally, all accepted hypotheses are combined to produce a final outcome. The detection model is found to perform better than human auditors. Deception tactics also frequently appear in warfare. In the work of Yuan (2007), the author proposed a detection method incorporated with an Adversary Intent Inference (AII) model in order to detect deception tactics in warfare. The AII model is used to infer the goals and actions of an agent given observed behaviors. After discrepancies are obtained by comparing the goals and the actions, deception branches that encode deception tactics according to the observed type of deception are attached to the original model in order to update the inference and

reduce the discrepancies. Different strategies are employed to detect simulative and dissimulative deceptions because of their different natures. This process is iterated until the goals and the actions converge. Another approach which prototypes a model combining many deception detection techniques was proposed by Vyas and Zhou (2005). The model covers a holistic detection process including searching for vulnerabilities and indications, analyzing logged information, and undoing the damages from deception. The intent of the deceiver and the environment are taken into consideration in order to collect more precise indicators. For example, potential deceptions can be indicated from specific vulnerabilities of the environment that may motivate the malicious intent of an agent and from any manipulation of the environmental information.

A third school of detection methods utilizes credibility or reputation of agents by querying their relationships with others during social interactions, pioneered by Schillo, et al. (2000). Their model of trustworthiness is built upon the agent's knowledge of the other agents' past behavior, honest or deceptive. The model may converge accurately after several rounds of decision making. However, the failure to catch the deceiver in the early rounds may result in irreversible loss. Moreover, this school of work is targeted at identifying deceptive sources by leveraging the experience of socially related entities but not by understanding the behavior or the communication of deceivers. In contrast to these approaches, our goal is to systematically study the intent of deceivers, and thoroughly analyze the observables of their intent.

The research that models the tactics of deceivers tackles the problem of deception detection in a top-down fashion, whereas studies measuring cues approach the problem in a bottom-up way. The former research provides a theoretical basis to the detection methods, but is usually domain specific and can only apply to datasets with restricted assumptions. Johnson et al.'s model is only applied to accountant reports; in Yuan's work, deception tactics are retrieved from a library of strategy fragments; Vyas and Zhou assume that entities in the environment have a conflict of interest and that the entities have up-to-date knowledge about the environment and each other. The later research evaluates verbal cues from real world datasets through trial and error without linking the findings to an overarching framework. In this dissertation, we attempt to approach the problem in both ways. Linking theoretical models of detection with real world data is an essential process of the study because on the one hand the theories can guide and support the exploration of empirical analysis, and on the other hand the empirical observations can evaluate the assumptions and hypotheses proposed in the theoretic model. In addition, we claim that by studying the reasoning of deceivers we can solve the problems that conventional methods are faced with in a radical way because the reasoning process of a speaker is directly driven by his intent. We want to develop a generic framework that can model the reasoning process of deceivers and then apply it to datasets of different topic domains.

1.5 Goal, Novelty and Contribution of this Dissertation

The purpose of this dissertation is to understand and quantify the reasoning of deceivers and provide a computational method of detecting deceptive reasoning with improved performance and reliability without losing the generality of modeling human intent. Our intuition is that a deceiver's goal of communication is to provide strong arguments for a falsified conclusion, whereas a truth teller's goal is to derive a conclusion based on arguments. In order to reach their communicative goals, deceivers need to presuppose their falsified conclusions before they form arguments. Due to this presupposition, deceptive arguments can be identified by looking for a speaker's deviation from himself and his discrepancy with truth tellers. Deviation from the self determines whether the speaker is acting like himself or not, whereas discrepancy with truth tellers determines whether the speaker is convincing or not. Violation of the two is an essential behavior of a deceiver. Since the presupposition of a falsified conclusion is an inevitable process in deception the reasoning patterns become the fundamental difference between deception and truth, and the identification of unique patterns becomes an accurate and reliable method of deception detection.

The novelty of this work is that it proposes an over-arching framework of deception detection independent of domain knowledge, performs empirical analysis based on observed patterns of deception, and connects the theoretic framework with the observational experience in such a way that the observations can be explained by the framework and the framework derives hypotheses of observation. We will show in the experiments that linking the theoretic framework with empirical studies is not a

straightforward task since many of the assumptions in the theory do not apply to the real world, and the noise and imprecision of real world data needs to be handled carefully. However, thanks to this connection, our work also bridges the gap between the qualitative evaluations and the quantitative measurement of attributes in communication content such as the convincingness of a text by retrieving and modeling the semantics of communication content.

To accomplish these goals, we propose to:

- Develop a domain-independent model that can capture the reasoning process of deceivers by realizing and enriching theories in cognitive science through computational methods. More specifically, we take advantage of the ideas in the theories of argumentation and the advanced techniques of knowledge retrieval and knowledge representation in AI.
- Detect deception through identifying and measuring unique reasoning patterns of deceivers. This method is threefold: it effectively improves the performance and reliability of deception detection, it explains deceptive communication in terms of the reasoning process of deceivers, and it guides the future development of cueing methods by finding the correspondence between the analytical results generated by the detection and the cues in the observational research.

The contributions of this work include the following:

1. Provide a domain-independent model of intent-driven reasoning.
    a. Propose a cognitive model of deceptive and honest reasoning.
    b. Propose, implement and analyze methods to retrieve semantics from natural-language stories.
    c. Represent and infer human knowledge using computational methods
2. Provide a computational method of deception detection using human reasoning
    a. Propose and implement a framework of deception detection
    b. Explain deception in terms of deceiver strategy and detection process
    c. Quantify and verify reasoning patterns of deceiver
    d. Analyze sensitivity of detection to parameters of datasets and speakers

This dissertation is organized as follows: In Chapter 2, we explain our intuition for using reasoning to detect verbal deception, which is followed by the description of our framework in Chapter 3. We also illustrate the application of our model using an easy to understand example. In Chapter 4, we discuss the performance of our model when applied to real world datasets, followed by a description of the techniques we use to process real data. We then discuss some observations from our analytical study. We finish up with a discussion of conclusions and future works in Chapter 5.

Chapter 2 From intent to reasoning

2.1 Overview

To detect deception effectively we want to look for the intent of deception and its direct product, but, why is intent so important in deception detection? First of all, we argue that deception is driven by malicious intent. Malicious actions are persistent. They do not only bring about damage by transmitting wrong information, but also purposefully aim at vulnerable targets over a long period of time until the targets' actions are manipulated. Since the malicious intent is as critical as, if not more critical than, the error in the information, there is a need to disclose the intent of the communicator so as to discriminate deceptive communication from all other unintentional errors. Intent is the key to deception detection. The main focus of the existing computational methods is on the word-level features of deceivers. However, thinking beyond the word level, we can trace the beginnings of a deceptive act to its malicious intent, which is to persuade the listeners of the truth of forged information. This intent does not exist in any kind of non-deceptive acts no matter whether they transmit true information or false information, thus it has the discriminative power to identify deception from both the unintentional errors that transmit wrong information and the intentional persuasions that argue according to a presupposed true conclusion. The identification of intent seems to be difficult since at first glance intent is too subtle to be captured computationally, but this is not true. The intent or purpose of a communication determines how people derive arguments before the communication. For example, thinkers derive conclusions based on evidence and persuaders produce valid arguments in order to support their conclusions (Walton, 2005).

Different types of speakers' unique reasoning processes generate unique patterns in their arguments. As long as we are able to find the patterns and measure them to some extent, we will be able to tell one reasoning process driven by a certain type of intent from another.

2.2 What is the advantage of using reasoning to detect deception?

The reasoning process of a communication is a cognitive process. To analyze the cognitive process of a deceiver in asynchronous formal communication, we studied how humans derive their arguments since formal communication is closely related to argumentation. Major studies that explore the cognitive process in argumentation includes Milkowski, 2008; Carenini and Moore, 2006; Zukerman et. al, 1998. These studies proposed architectures of argument generation given users' goals and preferences. The generation of arguments involves a system to generate the semantic structure of arguments and a system to convert the semantic structure into a human understandable language. By summing up these researchers' ideas, we conclude that the process of argumentation involves two stages: argument formation, and argument implementation. What a speaker does in the stage of argument formation is to infer the arguments based on his context knowledge and select appropriate arguments based on his preference and the communication requirement (such as the length of the story). What he does in the stage of argument implementation is to transform the selected arguments into a language output. The stage of argument implementation has been the focus of the research on verbal cues. Deception detection in the implementation stage is not ideal because as

artifacts of argumentation, a language output is mediated by different human and environmental factors and is easily to be manipulated. On the contrary, distorting arguments in the formation stage is extremely difficult because argument formation directly follows the intent and changes to argument formation may result in the failure of the intended goal. Also, most speakers are not aware of the process of argument formation as they communicate. Lastly, hiding deceptive arguments requires higher-order deception that takes the interlocutor's intent and even the interlocutor's belief about the speaker's intent into consideration. Higher-order deception demands much more cognitive load than first-order deception in order to retrieve the memory about the interlocutor's intent and leverage the original reasoning process behind it. Due to its high complexity and rarity, higher-order deception is usually not the focus of research in the deception detection community, nor within the scope of deception that we are interested in as discussed in Chapter 1.3. Thus, the argument formation stage provides more accurate and reliable observations than the argument implementation stage.

By switching our focus from the presentation of arguments to the formation of arguments, we realize that deception is not only a falsification of truth, but also a deviation from the original beliefs. This finding has been generally accepted in philosophy. For example, Shibles (1988) states: "A statement itself cannot be a lie. A lie is a relationship between two statements. An assertion must be compared to one's belief statement in order to determine if it is a lie… A lie is merely a contradiction between belief (self-talk) and expression." Therefore, the purpose of deception detection is not to find the most

irrational person but to identify the people who are contradicting themselves. Our prior work (Santos et. al, 2008; Santos and Li, 2010) already demonstrated that inconsistency within the speaker himself is a critical indication of deception, which can be precisely captured through the techniques adopted in recommendation systems. It is also not difficult to see that although cues of deception are innumerable and seemingly independent of each other, most of them can be related and explained by the way that deceivers generate their arguments. Without linking to argument formation, conventional research has failed to generalize their findings and explain exceptions. On the other hand, by considering the cognitive process of deceivers it is easier to not only detect deception in a robust way but also explain patterns for deceptive reasoning such as the effectiveness of arguments to reach the conclusion, as we have suggested in prior work (Li and Santos, 2012). Some of our findings have been observed from word-level cues, but none of the word-level cues were driven by the understanding of human reasoning. Therefore, to address the problems faced by existing technologies, researchers need to fill in the gap between the intent to deceive and the implementation of deception. Unfortunately, very little research can be found on the cognitive process of deceivers. Existing work generally neglects the fact that deception is rooted in the formation of arguments mainly because such process is not directly observable. This motivates us to derive a generic cognitive model to quantify the reasoning process of deceivers.

2.3 Cognitive Process in Argumentation

Deception detection cannot be accomplished without an understanding of truth tellers, hence we study the cognitive process in valid argumentation first. The task of argument formation is accomplished by a knowledge base which stores knowledge and a reasoning engine which processes knowledge.

2.3.1 Knowledge Base

The knowledge base contains all knowledge that is relevant to a topic. It can also be regarded as the context knowledge of the speaker with respect to the topic. The importance of context knowledge in argumentation has been pointed out by Carenini, (2006) and Zukerman et. al (1998). In (Carenini, 2006), context knowledge is represented as user models which provide the compellingness of different arguments in a user's perspective. In (Zukerman et. al, 1998), context knowledge is supplied by reasoning agents to provide knowledge that is relevant to the current focus. The context knowledge in a knowledge base is the subjective beliefs of the speaker. They may or may not be explicitly expressed by the speaker, but the speaker considers them relevant and makes inference through them during argument formation. Researchers (Falappa et. al, 2009; Pasquier et. al, 2006) propose that a knowledge base is structured as "a set of interrelated pieces of knowledge supporting the objective from evidence" and "interaction between arguments". Hence, a knowledge base can be structured as arguments with connections. Meanwhile, researchers like Zukerman et. al (1998), suggest the use of graphical representations in encoding arguments in which nodes represent propositional arguments and links represent inferences that connect the arguments. A

prototype knowledge base can be seen in Fig. 1. The nodes in the center circle denote the arguments explicitly mentioned, those in the middle circle are relevant to the arguments in the center circle, which may or may not be mentioned, and the relevance of nodes decreases as we move out of the center. In research on context knowledge this is called the "onion metaphor". Specifically, the knowledge is propositional arguments that can be assigned true or false with different levels of confidence. The knowledge can also be categorized into conclusion and support. Arguments are connected with other arguments on which they have direct impact such as causality, correlation and conditional relationship. Bearing these requirements in mind, we use Bayesian Networks (BNs) (Pearl, 1988) to represent the knowledge base of a speaker, in which the nodes refer to the arguments, the links refer to the relationship between arguments, and the conditional probabilities refer to the strength of the impact of the arguments on others. A short description of Bayesian Networks can be found in Appendix F. As suggested by Zukerman et. al (1998), the argument structure can be represented by BNs because of their ability to represent relationship between arguments and to normatively correct reason under uncertainty. BNs have already been used to make sense of both rational and emotional human reasoning during argumentation (Carofiglio and de Rosis, 2003) as well as for other purposes (Tenenbau et. al, 2006). BNs' reasoning schemes have been verified to show behavior similar to human beings (Tenenbau et. al, 2006).

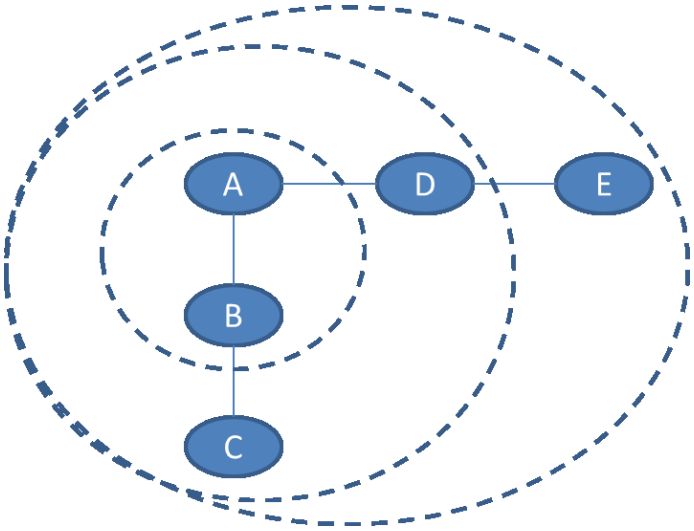


Figure 1 A prototype of knowledge base/ context knowledge

2.3.2 Reasoning Engine

The reasoning engine generates possible arguments that are constrained by the knowledge base and the evidence. The generation of arguments is governed "by some kind of rules, using standard or non-standard derivation to implement reasoning", according to Falappa et. al (2009). Levi (1996) argues that, during the process of reasoning, "the agent should be in suspense concerning which of the available options he or she will choose. If the agent is convinced that he or she would not choose a given policy, choosing and implementing are not serious possibilities according to that agent's state of full belief and the policy is not optional for that agent. The deliberating agent needs to be able to assess the implications of choosing the policy even though he or she is in doubt as to whether it will be chosen and implemented." This statement clearly illustrates the entire process of reasoning, which includes identifying existing beliefs, assessing possibilities of options given existing beliefs, and choosing the ones that are significantly more probable than the mutually exclusive others.

Belief updating, which infers the posterior probabilities of nodes based on Bayes' theorem, is not feasible for representing the reasoning in argumentation since it does not consider consistency within all arguments. Consider the following example: Suppose 10 people joined a lottery and exactly 1 of them will win. By belief updating, the result is that no one will win because all of them have a probability of 0.1. To retain the validity of the probability of each variable as well as maintain the consistency over all variables, we propose the following inference: We first perform belief revision and obtain the most probable explanation (MPE), which is the complete collection of nodes assigned with the states that can maximize the joint probability as in:

$$argmax_X p(X, E)$$

where X is the complete set of nodes, E is the set of evidence, and $p(X,E)$ is the joint probability of X and E. Then, for each variable we compute its posterior probability given that all other variables are set as evidence with the same assignment of states as in the MPE:

$$p(x_i | X - x_i, E)$$

where x_i is the i^{th} variable, and X- x_i is the complete set of nodes except x_i. By representing the lottery example in this way, in each of its inferred explanations, a different person wins with equal probability. Specifically, the probability of a person winning given all others not winning is 1, and the probability of a person winning given all but one winning is 0.

2.3.3 Argument Selection

According to Carenini and Moore (2006), usually for the sake of brevity a story cannot mention all of the available arguments. Only strong arguments are presented in detail, whereas weak arguments should be either briefly mentioned or omitted entirely (The weak arguments may be context knowledge because they are still relevant.) Thus, after the arguments are generated in the knowledge system, they are usually selected based on their strength. Earlier research has proved that the strength of an argument is judged on its plausibility in the decision maker's mind rather than on truth (Mann, 2010). Arguments and their strengths, from the view of Carenini and Moore (2006), should be determined according to the reader's values and preferences as suggested by argumentation theory. More specifically, they proposed that the strength of an argument is determined by its compellingness which is the product of its value and its importance to the objective of the speaker. Compellingness measures the argument's strength in determining the overall value difference between the alternatives of the objective, all other things being equal. An argument is notably-compelling (worth mentioning), if it is an outlier in a population of arguments with respect to compellingness. Carenini and Moore's (2006) measure of compellingness is a computational version of the argumentation theory. We borrow their idea and further generalize it to any type of argumentation.

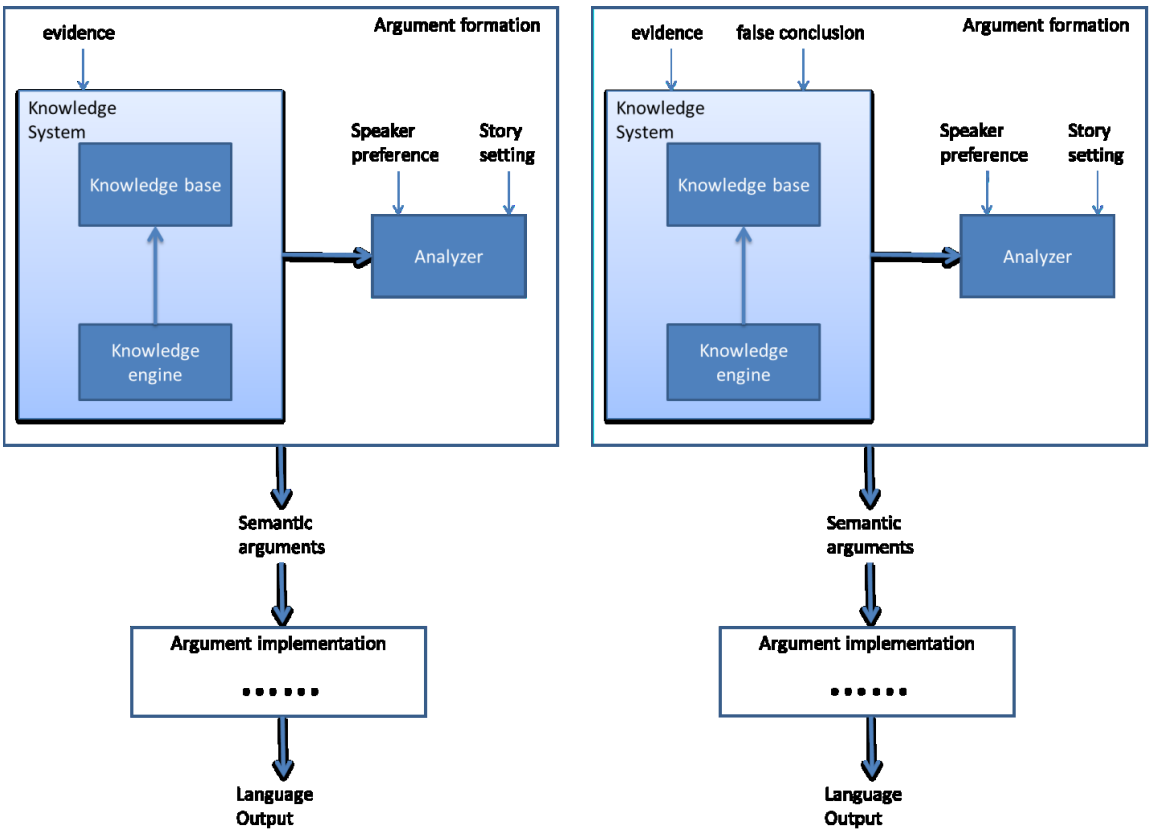
In a BN, the value of an argument is represented by its probability, alternative objectives of the speaker refer to the mutually exclusive states of a target argument, and an argument's importance to the target argument can be represented by their conditional dependence. An argument is compelling if its exact impact on the target argument is

significant. We can use the following formula to calculate the compellingness of an argument:

$$compellingness(A, B) = p(A) * [p(B|A) - p(\bar{B}|A)]$$

where $compellingness(A, B)$ denotes the compellingness of argument A in determining the target argument B, $p(A)$ denotes the probability of A, and $p(B|A)$ denotes the probability of B conditioning on A, $\bar{B}$ denotes the mutually exclusive state of B. An argument is notably-compelling (worth mentioning) if it is more compelling than other arguments in determining a target argument. Notably-compellingness is a decision criterion for including an argument in the story. Assuming that the compellingness of arguments in determining a target argument has normal distribution, we can decide that an argument outperforms the other arguments if its compellingness positively deviates from the other arguments' compellingenss by *k* standard (std.) deviations. Notice that the value *k* in determining notably-compellingness is a lower bound of compellingess for an argument to be included in the story. By setting *k* to different numbers, we can control the size of a story by including different numbers of arguments. Since the strength of arguments should be determined according to individuals' subjective preferences (Mann, 2010), the value *k* can be used to represent a speaker's personal tendency to believe in an argument. A speaker tends to believe in arguments if arguments with low compellingness are judged as worth-mentioning. The difference in the threshold also indicates that an ideal deception detection scheme should consider individual difference in understanding truth. If two people have exactly the same degree of belief but different compellingness thresholds, by applying universal heuristics, one may be regarded as hiding the truth or the other as exaggerating the truth while both of them may be honest.

The cognitive process behind argument formation can be depicted as in Fig. 2(a).



(a) (b)

Figure 2 (a) Truth teller's cognitive process behind argument formation; (b) Deceiver's cognitive process behind argument formation

2.4 Cognitive process in deception

With a better understanding of the cognitive process in the argumentation of truth tellers, we can start discussing the cognition of deceivers. Deception is a special type of argumentation with intentionally conveyed false information, but what type of argumentation is deception? Deception has been identified as a "goal-driven, intentional act" (Buller and Burgoon, 1996), which means that deceivers communicate in order to persuade listeners of the target conclusion (the objective) instead of reaching some other

conclusion. Hence we derive that deceivers possess a pre-targeted conclusion and argue according to it. It leads to a proposal that the act of deceiving is to derive statements by supposing the validity of the deceiver's claim which is actually not believed by the deceiver. For example, if a person is asked to lie about his attitude towards abortion, he might raise arguments such as "fetuses are human", "god will punish anyone who aborts children" and "children have the right to live". The reason he raised these arguments is not because he believed in them but because they support the false conclusion that he is against abortion. It is thus natural to imagine that the conclusion comes into the deceivers' minds before the arguments do. According to Levi (1996), "The addition of the supposition to the agent's state of full belief does not require jettisoning any convictions already fully believed. The result of this modification of the state of full belief by supposition is a new potential state of full belief containing the consequences of the supposition added and the initial state of full belief", which means that the reasoning with suppositions is regular reasoning with the addition of knowledge that has been assumed before the reasoning starts. The argument does not refute the possibility that the reasoning with a supposition may infer exactly the same arguments as regular reasoning in which the supposition in the former case is a true belief. In this case, deception detection becomes extremely difficult. However, the increase of reasoning complexity in the knowledge can reveal the difference between a deceiver's beliefs and a truth teller's beliefs by leaking cues on the arguments that deceivers missed to manipulate. As a result, it is less likely for a deceiver to generate the same arguments as a truth teller. Moreover, deceivers and truth tellers differ in their communicative goals. As a deceiver, one provides strong arguments to support the falsified conclusion, while as a truth-teller, one

provides arguments to infer the most likely conclusion. As a result, their beliefs and processes of reasoning are different even though they may claim the same conclusion. In particular, if an opinion-based story is required from the speaker, truth tellers propagate beliefs from evidence, while deceivers adapt beliefs to suppositions. If an event-based story is required, truth tellers retrieve relevant memory which is related to past behavior and past behavior is based on past belief, which was propagated from past evidence, while deceivers assume a part of the event and adapt this fantasy to the supposition. All in all, deception involves the intentional formation of arguments based on false beliefs, while truth involves the intentional or unintentional formation of arguments based on true beliefs. There are two cognitive processes that distinguish deception from legitimate argumentation. Firstly, the conclusion that is mutually exclusive to the original conclusion is fed into the knowledge system before reasoning. Secondly, the reasoning engine takes the false conclusion as an assumption in addition to observed evidence. Bearing in mind that a rational person's reasoning resembles Bayesian knowledge inference, we can simulate a deceiver's reasoning by setting a wrong state of his conclusion as evidence. The cognitive process behind deceptive argument formation can be depicted as in Fig. 2(b). This fundamental difference in the reasoning of deceivers and truth tellers is unavoidable due to the intentionality of deceivers. It provides a theoretic basis according to which schemes of deception detection can be built.

Our selection to utilize BNs and the theory of argument compellingness to represent the knowledge system is one of many possible ways to interpret human reasoning. Our proposal of the knowledge system guides the development of the detection model which

will be presented in the next chapter. Our detection method assumes that the arguments are related in some way, but it can be applied independent of the representation of human reasoning. The assumptions in this chapter are only critical to the learning of speaker's cognitive process. The learning of cognitive processes is a complex task as human reasoning is featured with incompleteness, uncertainty and ambiguity. We encountered these challenges when dealing with real data, which will be discussed in Chapter 4.3.

Chapter 3 Deception Detection using Human Reasoning

3.1 Overview

Through the exploration of the cognitive models, it is clear that deception starts from a malicious intent to manipulate others, which then drives the reasoning with false presuppositions. Although we cannot see reasoning directly from communication output, we can obtain the results of reasoning, which are the semantic arguments. Since the semantic arguments are determined by the reasoning process, the type of reasoning, deceptive or honest, is embedded in the semantic arguments in some fashion. Therefore, the task of deception detection, or detection of any type of reasoning, is to find observable patterns from the semantic arguments and their relationships.

3.2 Model of Detection

Inspired by Johnson et al.'s work (2001), we propose a model of deception detection composed of three modules: (i) discrepancy detection, (ii) reasoning pattern identification, and (iii) classification. Discrepancy detection finds suspicious deviations by comparing observations with expectations. Reasoning pattern identification attempts to explain the unexpected deviations with respect to how strong they fit the hypothesized patterns of deception. Their materiality in each pattern is combined in the classification module in order to classify the input text to deception/truth.

3.2.1 Module I: Discrepancy Detection

What kind of reasoning patterns are the most effective for deception detection? Probably the ones that frequently appear in either deceptive reasoning or honest reasoning but not both can best serve for deception detection. As we have discussed, truth tellers and deceivers may share the same reasoning results. However, they do not usually derive the same arguments in the real world because they do not share the same belief system that supports their reasoning. If, in the case they do share the same belief system, they would likely reach the same conclusion without any deception and thus, there would be no need to deceive. Deceivers may be successful in manipulating conclusions in a way to mimic truth tellers' conclusions and distorting arguments to support the manipulated conclusions, but the supporting arguments are biased by their original beliefs. Psychological studies that echo this finding (Johnson and Raye, 1981; Markus, 1977; Mehrabian, 1972, Wiener and Mehrabian, 1968) suggest that deceivers are not able to produce the same stories as truth tellers because (i) their stories are solely based on imaginations, and they do not have the same knowledge to back their stories up, and (ii) they do not embrace the stories as much as truth tellers do since they do not believe in the stories. These two points clearly summarized the difference between deceivers and truth tellers in terms of reasoning. To put them in another way, a deceiver's inferred arguments are different from truth tellers' as well as from his own beliefs. That is to say, by comparing a deceiver with himself and with the truth tellers we are able to identify potential deceivers through expecting the following discrepancies in a deceptive story:

1. Discrepancies in arguments that are manipulated by deceivers can be expected because (i) arguments with presupposition may be different from arguments

without presupposition, and (ii) arguments inferred from false beliefs are different from arguments inferred from true beliefs.

2. Discrepancies in arguments that deceivers are reluctant to believe but truth tellers embrace can be expected because inferences based on knowledge that cannot reach the claimed conclusion is different from inferences based on knowledge that can reach the claimed conclusion.

We will name the first discrepancies as "inconsistency" because behaving differently from the self indicates inconsistency, and the second discrepancies as "untruthfulness" because it deviates from the truth provided by others. Our focus is to explain and measure them in terms of human reasoning. Since these two principles of discrepancies agree with the studies in various domains (Mehrabian, 1972; Wiener and Mehrabian, 1968; Johnson and Raye, 1981; Markus, 1977), it suggests that our proposal of the deceptive reasoning is a reasonable explanation of the cognitive process in deception.

Although untruthfulness is straightforward to measure, the measurement of inconsistency is complicated because individuals' own beliefs are subjective and hidden. This requires us to avoid implementing general rules but customize rules of detection according to each individual. In our daily life, it is not difficult for anyone to predict/anticipate the behavior of close acquaintances. Although people's behavior may not be perfectly logical or rational, they are predictable because someone knows them so well. Tindale (1999) also agrees that "The rational agent is more predictable than many other human characters." This intuition can be applied to our method for the purpose of inconsistency detection.

We can expect that honest people are consistent over time. Even if a person may be different from others, he is consistently different. Thus, we still can predict his opinion based on opinions from other people. If his exact opinion deviates too much from what we expect, his opinions are regarded as inconsistent. This is more realistic than to expect people to agree with each other because distinctive opinions do not make a person deceptive but might indicate innovation and make him more valuable to the group. Computationally, it means that given a pair of agents whose past arguments correlate with each other, it is possible to calculate the expected value of one agent's future arguments given the other agent's future arguments. The mathematical method was first proposed by GroupLens and applied to recommendation systems (Resnick et al., 1994). It was later applied to deception detection by Santos and Johnson (2004) and discussed in our prior work (Santos and Li, 2010) in detail. We generalize and incorporate the method in our model.

The measurements of inconsistency and untruthfulness give us an indicator of unexpected deviations in one's stories. They can be obtained from the following framework (Fig. 3). The first module (Discrepancy Detection) of the framework is composed of two networks: the *correlation network* and the *consensus network*. The correlation network connects each speaker with others who correlate with him in a specific argument. Neighbors in the correlation network represent acquaintances who can anticipate each other's arguments. The consensus network connects speakers with similar conclusions. Neighbors in the consensus network represent people who agree with each other. We have pointed out that deception is deviations from one's own subjective beliefs, but not

deviations from the objective reality or from the public. Thus, the correlation network is essential in predicting a speaker's belief according to the neighbors who can anticipate his opinions, and the prediction can be used to evaluate the speaker's inconsistency. The methodology of inconsistency detection can be achieved using the following steps:

1) Compare Opinions: An acquaintance's opinion can be anticipated by correlating one's own opinion with him. This observation enables us to predict one's opinion based on his correlation with others. Thus, the first step is to calculate the correlation between each two speakers based on multiple repeats of their past opinions. The historical reasoning process is also called the *training process*, and the opinions generated in the past are called the *training data*. We assume that the training data does not contain any deceptive opinion. Thus, it does not play a role in identifying inconsistency but is used to obtain the correlation values. The correlation measure we use is the Pearson Correlation, which is calculated in the following formula:

$$r_{AB} = \frac{Cov(A,B)}{\sigma_A \sigma_B} = \frac{\sum_i (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_i (Ai - \bar{A})^2 \sum_i (B - \bar{B})^2}}$$

where r_{AB} represents the Pearson correlation coefficient. For the i^{th} repeat, we define A_i as the posterior probability of speaker A, and B_i as the posterior probability of speaker B. $\bar{A}$ denotes the average of the probabilities assigned to speaker A over all repeats, and likewise for B.

2) Predict Opinions: After the correlations are calculated, we predict each speaker's opinion in a repeat using the opinions of his acquaintances, where a speaker is an

acquaintance of another speaker if their correlation is beyond a threshold. The reasoning process from which we want to detect inconsistency is called the *testing process*, and the opinions generated from the reasoning process are called the *testing data*. The technique we use to predict opinion is based on GroupLens prediction, which allows us to estimate what opinion is expected for each speaker provided that his historical opinions are sufficient:

$$A^t_{j_{prediction}} = \bar{A_j} + \frac{\sum_{i \epsilon \mathbb{R}_A} (A^t_i - \bar{A_i}) r_{A_i A_j}}{\sum_{i \epsilon \mathbb{R}_A} |r_{A_i A_j}|}$$

where $A^t_{j_{prediction}}$ denotes the predicted probability of the j^{th} speaker A_j in repeat t, $r_{A_i A_j}$ is defined as the Pearson correlation coefficient between A_i and A_j, and $\mathbb{R}_A$ represents the correlation network with the set of speakers A. $\forall i, j < N, i, j \in \mathbb{R}_A$ iff $|r_{A_i A_j}| > \Delta$ where N is the total number of speakers.

3) Identify Inconsistency: If a speaker's actual opinion on a given problem is very different from the predicted opinion, it means that he provided an inconsistent opinion. In practice, we predict the benevolent training data, measure the errors of prediction. and use the prediction errors as a reference for legitimate noise. We assume that the prediction error has a normal distribution. If the prediction error of the testing data deviates from the prediction errors of the training data by more than some std. deviations, the testing data is regarded as inconsistent. Inconsistency is measured as how many number of std. deviations a speaker's prediction error deviates from his historical prediction errors.

The consensus network provides a sampled population of truth tellers who reach the same conclusion as the deceiver. If the deceiver had told the truth, he should have behaved no differently from the population. The deviation from the truth tellers can be measured using std. deviations:

$$\sigma_A^t = \sqrt{\frac{1}{|\mathbb{C}_A^t|}\sum_{i\in\mathbb{C}_A^t}(A_i^t - \overline{A^t})^2}$$

where σ_A^t denotes the std. deviation of the population A in repeat t, $\overline{A^t}$ denotes the average of the probabilities assigned to all speakers in A in repeat t, and $\mathbb{C}_A^t$ represents the consensus network with A given the same set of evidence in repeat t. $\forall i, j < N, i, j \in \mathbb{C}_A^t$ iff A_i and A_j have the same conclusion in repeat t. If the probability assigned to a speaker in $\mathbb{C}_A^t$ deviates from the population by more than some numbers of σ_A^t, it is regarded as untruthful. Likewise, the untruthfulness is measured as how many number of std. deviations a speaker deviates from the population in the consensus network. The correlation network explains why the deceiver is convincing, or what manipulations make his story convincing, whereas the consensus network explains why the deceiver is unconvincing when compared against the real truth tellers.

To summarize, the first module of the framework detects the basic discrepancies in deceptive reasoning, which are inconsistency and untruthfulness. Inconsistency means that the arguments in the story contradict with what the speaker believes. It is particularly important when individual difference is considered. Untruthfulness means that the arguments in the story contradict with what a truth teller believes in order to reach the conclusion. Measuring untruthfulness is particularly effective in detecting deception from

strangers. On the other hand, the principle of inconsistency indicates that an honest person should behave as he always does, the detection of which requires some familiarity with the speaker, whereas the principle of untruthfulness indicates that an honest person should behave as a reasonable and convincing person, the detection of which requires some knowledge of the topic domain.

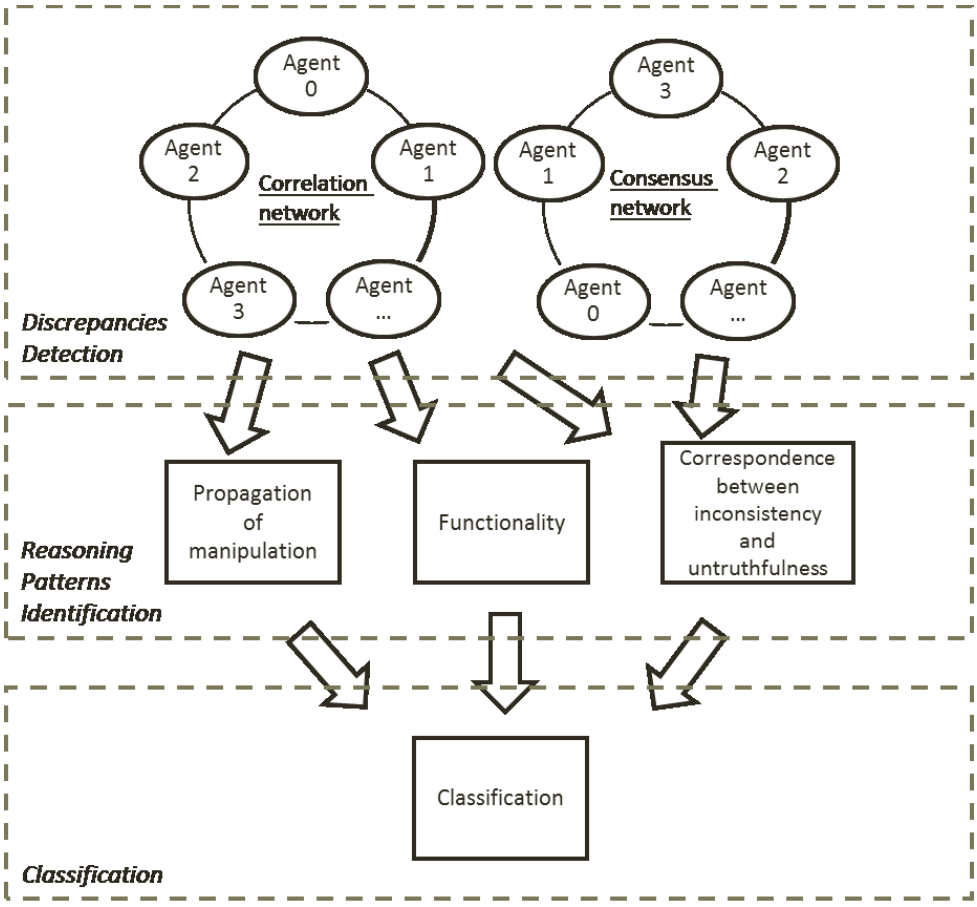


Figure 3 Architecture of the model of deception detection

3.2.2 Module II: Reasoning Pattern Identification

Deceivers do not deviate from themselves and from truth tellers in random ways because manipulations without the purpose to deceive also exhibit inconsistency and/or untruthfulness. For example, a changed opinion contains inconsistency but not

untruthfulness as it changes the prior knowledge while still maintaining to be truthful, whereas innovation may not conform to traditional attitudes but is still consistent as innovation is unpredictable truth. Yet, other types of misinformed opinion may exhibit both discrepancies such as misunderstanding, but obviously they are not derived from deceptive reasoning, and thus are expected to show different patterns of deviation.

3.2.2.1 A Case Study

We propose that deceivers can be distinguished by the manner they manipulate arguments. To see how a deceiver's reasoning results differ from an honest teller's, we take the example of a lawsuit case shown in a TV drama in which a female celebrity deceptively claimed that she was raped by a young Indian male. The deceiver's arguments sound convincing at first, but were later found to be deceptive by the jury. In the study, we perform two tasks: model the reasoning process of the deceiver and that of a hypothetical truth teller following our proposed cognitive models in order to verify the capability to identify discrepancies, and simulate unintentional misinformation so as to explore unique patterns in a deceiver's reasoning results. The detail of the case is described as follows:

> *A female celebrity coded as A claims that she was raped by an young Indian male, coded as B. A claims that she keeps away from B because both her and her mother do not like the physical odor of Indians. A claims that B once joined her birthday party without any invitation and fed A drugs. B then conveyed A home and raped A. After A's boyfriend arrived, A called police. However, the truth is that B is a fan of A and joined A's party*

at A's invitation. A lied about her aversion to Indians because she used to prostitute to Indians. Besides, B is new to the party club, so it is unlikely for him to obtain drugs there. A used drugs and enticed B to have sex with her.

This artificial scenario is a simplification of a possible legal case, which provides more realistic explanations as opposed to data that simulate deception arbitrarily without considering the intent of the deceiver. We did not choose to model real cases or surveys because it is difficult to find any survey with sufficient information about the reasoning of the deceiver, while data from real cases usually lacks ground truth. Data with both sufficient information and ground truth, such as military combat scenarios, is mostly focused on behavioral deception instead of communicative deception. In addition, real cases may contain noisy data in which the communication content is mediated by factors other than reasoning. For the purpose of exploring patterns in deceptive reasoning it is ideal to use clean data that only contains the semantic meaning of arguments. We use a BN to represent the knowledge base of *A* as discussed earlier. The semantic meaning of arguments and their conditional relationships are encoded in the BN. For example, the causal rule that *B* drives *A* home because *B* knows *A*'s address can be encoded in the conditional probability $P(B\_drive\_A\_home|B\_know\_A\_s\_adr)=0.9$. We designed a BN representing *A*'s belief system and another BN representing the belief system of a hypothetical victim of the rape case according to the description of the scenario. More specifically, we connect two arguments if their causal relationship is explicitly described by the deceiver or by the jury when they are analyzing the intent of the deceiver. The conditional probabilities between states of arguments are set as 0.7 to 0.99 according to

the certainty of the speaker if the arguments are explicitly described. As to the states that are not mentioned in the case, they are usually implied in or can be inferred from the scenario if their mutually exclusive states are described in the scenario, such as the probability of *A_hate_Indian* given that *B*'s relation with *A*'s mother is good and that *A* used to prostitute to Indians. Otherwise the mutually exclusive states of an argument are given the same or similar probabilities indicating that they are uncertain. To make sure that the discrepancies in deception are resulted from the manner of reasoning instead of from the inherent difference between the deceiver's belief system and the hypothetical victim's belief system, we minimize the difference between their belief systems. Specifically, we keep all their conditional probabilities the same by assuming that both *A* and the hypothetical victim are rational people with the same domain knowledge. Only their prior probabilities of *A*'s experience as prostitute and whether *B* is new to the party are adjusted differently because they are the essential truth from a victim's perspective. That is to say, those who do have strong prejudice against Indians could not prostitute to them, and to obtain drugs from the party club, *B* has to be a regular guest. As a result of sharing a similar belief system with the hypothetical victim, the deceiver's story may become highly convincing. Although we expect it to be hard to detect the untruthfulness of the deceiver, the deceiver's simulation is not unrealistic because some deceivers are consistently found to be more credible than others based on the study by Bond and Depaulo (2008). It is likely that a randomized BN with a perturbed copy can also serve the purposes of this study, but again, building belief systems based on the intent of deception will provide more realistic data, more convincing results and more intuitive

explanations. The BN of the deceiver is depicted in Fig. 4. The conditional probability tables can be found in Appendix D.

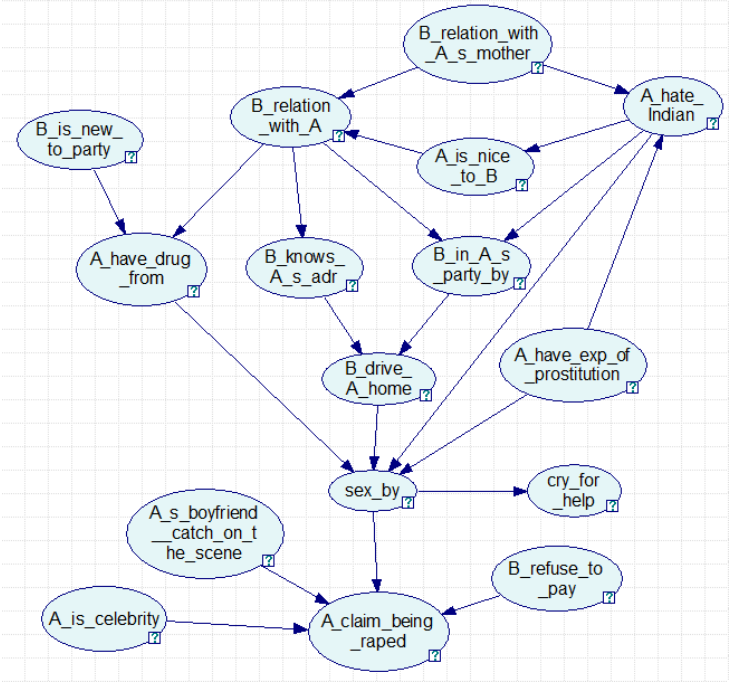


Figure 4 BN of the deceiver in the rape case

We use the inference scheme proposed in Chapter 2.3.2 to perform the reasoning, and the product of reasoning is represented by the inferred probabilities of the nodes. As we proposed earlier, in the reasoning process, the deceiver presupposes her target argument, that is, she was raped, by adding the argument as an additional piece of evidence. The inference results of *A* in both deceptive and honest cases and the inference results of the hypothetical victim are shown in Table 1. The arguments *B_relation_with_A_s_mother=bad*, *B_drive_A_home=true*, *A_is_celebrity=true* and *A_s_boyfriend_catch_on_the_scene=true* are set as evidence as suggested by the scenario.

People express arguments as binary beliefs (believe or not believe in something) in communication if not as beliefs with certainty modifiers, but not as degree of belief formulated by real-valued probabilities. To map degree of belief to binary beliefs, we need to know how strong an argument needs to be for a person to believe in it, or in other words, what is the probability threshold of something being true. Research has suggested that truth threshold varies by proposition and by individual, which means it is a subjective criterion (Ferreira, 2004). In Chapter 2.2.3, we proposed to use the theory of compellingness from Carenini and Moore (2006) to select strong arguments. In this study, as we use simulated data, we arbitrarily choose the probabilities 0.67 and 0.33 as the thresholds since they equally space the interval of an argument being true, unknown and false. By thresholding the probabilities of each argument to prefer the state with a high probability we generate the semantic arguments of the deceiver, the honest beliefs of the deceivers, and the beliefs of the hypothetical victim, as depicted in Table 1. For the purpose of exploration in this pilot study, we simplified the selection of explicit arguments, but we will incorporate the full scheme of argument selection in the experiments on real life data. To verify the inferred beliefs we compare Table 1 with the scenario. An argument is validated if it is in the same state as described in the scenario or in the unknown state if it is ignored in the scenario. We verified that 13 out of the 16 arguments in the deceptive story correspond with what the deceiver claims, and, all of the arguments in the honest story correspond with what is the truth. Although we cannot directly verify the hypothetical victim's story as we do not have the ground truth, we observe that all the arguments are reasonable and most of them, except the evidence, are contrary to the deceiver's honest story.

In looking for reasoning patterns, we should first detect the inconsistency and the untruthfulness of the deceiver. According to the detection model, the computation of the discrepancies assumes some familiarity with the deceiver, which requires a sufficient number of history data and acquaintances of the deceiver. To this end, we simulated 19 agents by perturbing the deceiver's BN and another 10 agents by perturbing the victim's BN. In total, we have 29 truth telling agents and 1 deceiving agent. We simulated 100 repeats of training data by inferring the network of each agent 100 times with a different set of evidence in each repeat, and convert the inferred probabilities to binary beliefs. We also observe patterns on an unintentional misinformed story in order to find out the patterns that are unique to deceivers. A misinformed story is simulated by adding random errors to the probabilities of the true arguments.

Table 1 Inferred results of the deceiver's deceptive story, her honest story and the hypothetical victim's story (left: inferred probabilities of arguments; right: selected states of arguments)

Arguments	Deceptive	Honest	True		Arguments	Deceptive	Honest	True
B_relation_with_As_mother=good	0	0	0		B_relation_with_As_mother	bad	bad	bad
A_have_exp_of_prostitution=T	0.66	0.88	0.11		A_have_exp_of_prostitution	unknn	T	F
A_hate_Indian=T	0.74	0.07	0.89		A_hate_Indian	T	F	T
A_is_nice_to_B=T	0.18	0.88	0.18		A_is_nice_to_B	F	T	F
B_relation_with_A=rape	0.98	0.16	0.96		B_relation_with_A	rape	fan	rape
B_in_A_s_party_by=self	0.9	0.4	0.90		B_in_A_s_party_by	self	unknn	self
B_knows_A_s_adr=T	0.95	0.95	0.95		B_knows_A_s_adr	T	T	T
B_drive_A_home=T	1	1	1		B_drive_A_home	T	T	T
B_is_new_to_party=T	0.76	0.82	0.16		B_is_new_to_party	T	T	F
A_have_drug_from=B	0.76	0.07	0.92		A_have_drug_from	B	self	B
sex_by=rape	0.93	0.08	0.98		sex_by	rape	entice	rape
As_boyfriend_catch_on_the_scene=T	1	1	1		As_boyfriend_catch_on_the_scene	T	T	T
A_is_celebrity=T	1	1	1		A_is_celebrity	T	T	T
B_refuse_to_pay=T	0.8	0.85	0.50		B_refuse_to_pay	T	T	unknn
A_claim_being_raped=T	0.6	0.7	0.60		A_claim_being_raped	unknn	T	unknn
cry_for_help=T	0.8	0.2	0.80		cry_for_help	T	F	T

We use the method described in Chapter 3.2.1 to detect the inconsistency within the stories. More specifically, we predict binary beliefs in the deceptive story using

GroupLens method (Resnick et. al, 1994) based on stories of neighboring agents in the correlation network. We then compare the binary beliefs in the deceptive story with predicted binary beliefs to measure the deviation of each argument due to inconsistency. We measured how many std. deviations the prediction error in the deceptive story deviates from the prediction errors in the training data, and plotted them according to their locations in the BN, which is shown in Fig. 5, together with the results from the misinformed story. The width of the links represents the sensitivity of each variable to its neighbors. The result from the deceptive story first verified that we are able to detect most of the manipulated arguments. If we compare the deceptive story with the honest story in Table 1, we obtain 9 arguments manipulated by the deceiver. Out of these 9 arguments, 8 are successfully identified as inconsistent as in Fig. 5 if we assume that the inconsistency threshold is 3 std. deviations. Next, we observe that the variables at the boundaries of the graph and not sensitive to neighbors (e.g. *B_is_new_to_party*) are ignored by the deceiver, while the variables in the center or sensitive to other inconsistent variables (e.g. *A_hate_Indian*) are manipulated significantly. It seems that manipulations propagate to closely related arguments. Unrelated arguments are probably considered as irrelevant or are simply ignored by the deceiver. However, in the misinformed story, only two arguments (*B_is_new_to_party* and *B_refuse_to_pay*) are found to be inconsistent. No evidence shows that the inconsistencies propagate to related nodes.

Paying more attention to the inconsistent nodes in the deceptive story (*A_have_exp_of_prostitution=unknn*, *A_hate_Indian=T*, *A_is_nice_to_B=F*, *B_relation_with_A=rape, B_in_A_s_party_by=self, A_claim_being_raped=unknn,*

cry_for_help=T), we found that all but one argument (*A_claim_being_raped=unknn*) strongly support the conclusion when compared with their honest states. Based on this finding, we hypothesize that the manipulated arguments are effective in reaching the goal and at the same time satisfing the evidence. This finding has also been brought up in cognitive studies and selected as verbal cues to detect deception (DePaulo et al., 2003). In this dissertation, we use "functionality" to refer to the effectiveness of an argument in supporting another. Being functional to the conclusion and the evidence indicates that an argument can be expected from the goal and the evidence. Hence we can estimate the functionality of each manipulated argument in the following way: For each inconsistent argument, we measure its correlations with the other arguments during the past using the training data. We then predict each argument's binary belief based on the value of the conclusion and the values of the evidence using the GroupLens method. If the predicted belief agrees with the belief in the deceptive story, the variable is regarded as functional. We compare the functionalities of the deceptive arguments with those of the misinformed arguments. As the results show in Table 2 and Table 3, all but one inconsistent argument in the deceptive story complies with the value expected by the conclusion and the evidence, but none of the inconsistent arguments in the misinformed story does. Although the result shown in Table 3 came from a random sample of the misinformed stories, we observed that most of the samples show the same functionality rate. Therefore, the functionality rate of the deceptive story is 6/7, and the functionality rate of the misinformed story is around 0/3.

Table 2 Functionality of the deceiver's story

Arguments	Pred.	Decept.
A_have_exp_of_prostitution=T	0.24	0.5
A_hate_Indian=T	0.85	1
A_is_nice_to_B=T	0.07	0
B_relation_with_A=rape	0.99	1
B_in_A_s_party_by=self	1	1
A_claim_being_raped=T	0.58	0.5
cry_for_help=T	0.86	1

Table 3 Functionality of the misinformed story

Arguments	Pred.	Misinfo.
B_in_A_s_party_by=self	0.45	0
B_knows_A_s_adr=T	0.90	0.5
A_claim_being_raped=T	0.94	0.5

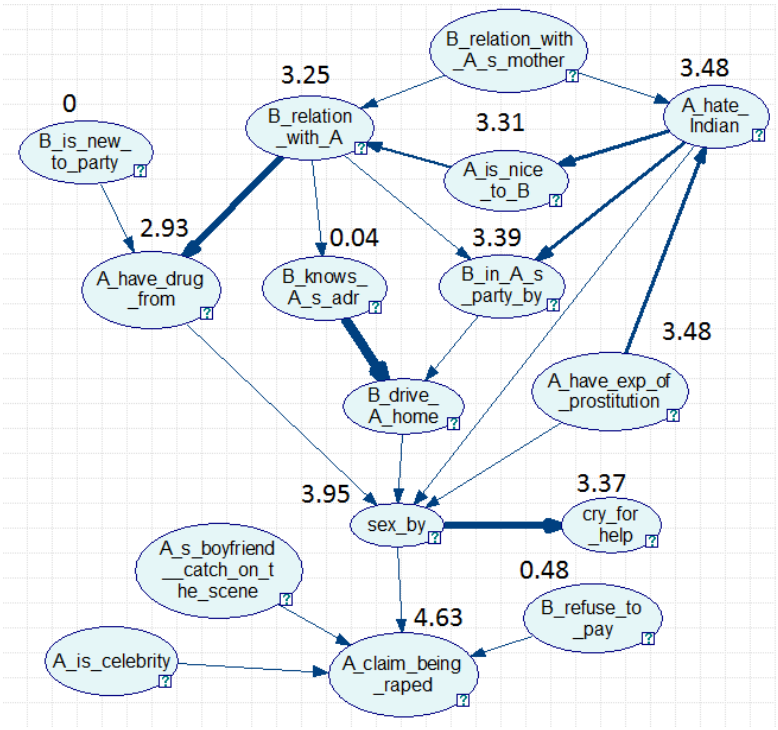
Next, we want to compare the deviations due to inconsistency with respect to the deceiver herself and the deviations due to untruthfulness with respect to the truth tellers. To compute the untruthfulness, we calculate the deviations of the binary beliefs in the deceptive story from the population of truth tellers' stories who agree with the deceiver in the consensus network using the method in Chapter 3.2.1. The results are shown in Table 4. If we compare the deceptive story with the victim's story in Table 1, we found that 3 arguments in the deceptive story are untruthful. The result of the untruthfulness shows that 2 of the 3 arguments are beyond 1.44 std. deviations of the population of true stories, and all of them are beyond 0.95 std. deviations. The small deviations indicate a high credibility of the deceiver, which is caused by the similarity between the belief systems of the deceiver and the victim. Among all 8 inconsistent arguments, none of them is identified as untruthful if we assume that the untruthfulness threshold is 1 std. deviation. It implies that significant manipulations are often convincing and unconvincing arguments usually can be found in slightly manipulated or ignored arguments. The only exception in the result is the argument *B_knows_As_address*, which is not manipulated

but convincing. It is probably because the evidence *B_drive_A_home* enforced it to remain honest. Compared with the deceptive story, all inconsistent arguments in the misinformed stories are regarded as untruthful, which shows that the negative association cannot be found in misinformed stories.

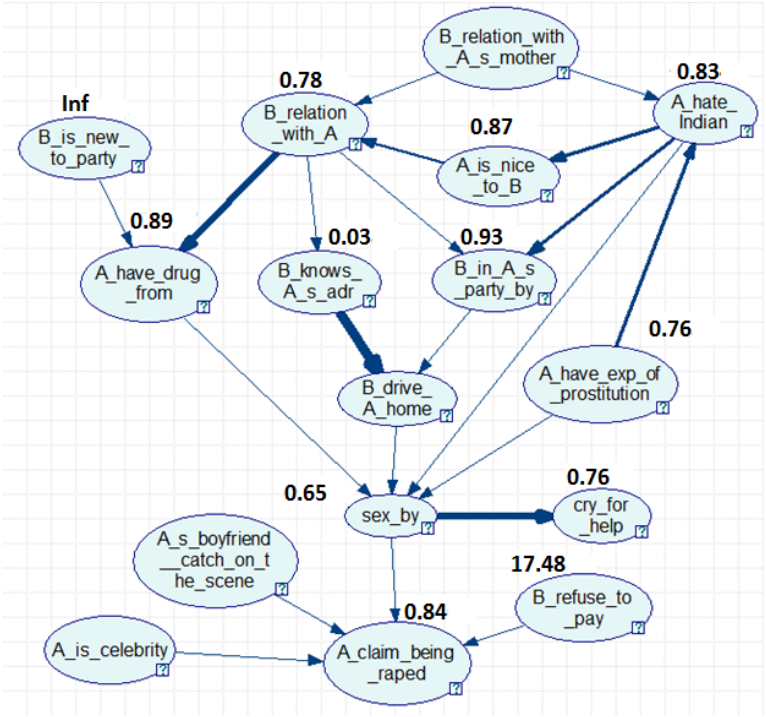
Table 4 Comparison of inconsistency and untruthfulness of the deceiver

Belief	Incon.	Untru.
B_relation_with_As_mother=good	N/A	N/A
A_have_exp_of_prostitution=T	3.48	0.95
A_hate_Indian=T	3.48	0.28
A_is_nice_to_B=T	3.31	0.28
B_relation_with_A=rape	3.25	0
B_in_A_s_party_by=self	3.39	0.28
B_knows_A_s_adr=T	0.04	0
B_drive_A_home=T	N/A	N/A
B_is_new_to_party=T	0	1.59
A_have_drug_from=B	2.93	0
sex_by=rape	3.95	0
As_boyfriend_catch_on_the_scene=T	N/A	N/A
A_is_celebrity=T	N/A	N/A
B_refuse_to_pay=T	0.48	1.44
A_claim_being_raped=T	4.63	0.41
cry_for_help=T	3.37	0.41

A summary of the computations performed in this section is listed in Table 5. From this study we discovered three patterns in deceptive reasoning: manipulations propagate through closely related arguments, deceptive arguments are usually functional to the deceiver's goal and evidence, and inconsistency and untruthfulness are negatively associated, which may serve as effective and measurable patterns to reveal deceptive intent given the arguments of a speaker.



(a)



(b)

Figure 5 Inconsistency deviation of each variable (a) of the deceiver's BN, and (b) of the misinformer's BN

Table 5 Computation steps of the pilot study on reasoning patterns

Step 1	Infer stories. (Table 1)	Perform reasoning on $Agent_D$ with evidence to get the value of the honest arguments.
		Perform reasoning on $Agent_D$ with evidence and presupposition of the false conclusion to get the value of the deceptive arguments.
		Perform reasoning on $Agent_V$ with evidence to get the value of the true arguments.
Step 2	Simulate misinformed stories.	Generate multiple copies of $Agent_V$'s story
		In each copy, perturb the probabilities by adding random noise
Step 3	Covert the values into binary polarities (Table 1)	
Step 4	Detect inconsistency and untruthfulness.	
Step 5	Evaluate the detection of inconsistency in the deceptive story	Compare inconsistent nodes with the difference between the deceptive arguments and the honest arguments
Step 6	Evaluate the detection of untruthfulness in the deceptive story	Compare untruthful nodes with the difference between the deceptive arguments and the true arguments
Step 7	Evaluate the propagation of manipulation in the deceptive story and the misinformed story (Figure 5)	Plot the inconsistency on the corresponding r.v.s
		Compare the inconsistencies with neighbors' inconsistencies
Step 8	Evaluate the functionality in the deceptive story and the misinformed story (Table 2 & Table 3)	Predict the value of each manipulated argument from conclusion and evidence using GroupLens
		Compare the predicted values with the actual values
Step 9	Evaluate the association of inconsistency and untruthfulness in the deceptive story and the misinformed story (Table 4)	Find out the proportion of inconsistent arguments that are untruthful

3.2.2.2 Reasoning Patterns

The case study provided an intuitive demonstration of the patterns that we can find from deceptive reasoning. It indicates that deception and truth are separable, but it is not clear why we are able to observe the patterns or how to quantify their boundaries in truth and deception. Therefore, we plan to find explanations of the patterns following the cognitive process of deceivers, and develop computational methods to quantify the hypothesized patterns in such a way that deceptive reasoning can be distinguished.

Stories told by experienced deceivers usually sound convincing. However, with only the prior knowledge and evidence, a deceiver's true beliefs are in conflict with what he wants to convey. Therefore, to make his arguments flow to the conclusion smoothly, he has to manipulate one or more arguments. A reasonable way is to let the distortion of the original conclusion impact related arguments, which can be better accomplished through natural reasoning instead of explicit manipulations of individual arguments. The exact impact depends on the network, but we can derive that the more closely an argument relates to the conclusion, the stronger it will be impacted. Likewise, the impacted argument will influence its related arguments. This is probably why manipulations can propagate to dependent arguments. If we imagine the structure of connected arguments as a terrain of arguments, the manipulation of the conclusion would transfer to the entire terrain like the epicenter of an earthquake, with related arguments receiving higher levels of manipulation and unrelated arguments receiving lower levels of manipulation. The purpose of propagating the manipulation is to derive convincing arguments to support the manipulated conclusion. Without it, the deception would be as blatant as insisting an illogical conclusion, which rarely happens in serious deceptions. Since the manipulations in misinformed stories are not generated by natural reasoning, misinformed stories do not exhibit this pattern.

The definition of deception indicates that deception is an intentional act. Referring back to Chapter 1.3, an intentional behavior must be driven by a desire, and to fulfill the desire a person adjusts his perception of the world to a yet to be realized state. In the context of deception, deceivers fit arguments to states that can support his targeted conclusion for

the desire of misleading listeners. They are persuaders, who reach arguments from conclusions, while others reach conclusions from arguments. According to the theory of satisfaction of intention (Mele, 1992), an intention is "satisfied" only if the behavior in which it issues is guided by the intention-embedded plan. This means that deceivers choose the best behavior (argument) that is guided (inferred) by his desire (conclusion), but not any behavior that can fulfill his desire. In particular, deceivers will choose the states of arguments in the story that are most effective compared with the other states of arguments in reaching the conclusion of the story. For example, in an honest communication, a speaker who claimed that he walked his dog today would equally likely say that the weather is good or that the weather is bad assuming that the listener has no knowledge about the weather. But by presupposing that he walked his dog while he actually did not, a deceiver is more likely to argue that the weather is good because it sounds more logical and supportive. Although there is nothing wrong if a truth teller's arguments are all functional to the conclusion, research has demonstrated that it is not a common conduct among truth tellers (DePaulo et al., 2003). In fact, truth tellers tend to make some arguments that conflict with their conclusions. On the other hand, in Bayesian belief revision, functional and nonfunctional nodes may result in the same MPE. Therefore, from both the psychological perspective and the computational perspective of reasoning, truth tellers' arguments are less functional than deceivers'. The functionality of an argument can be revealed from the history data of a speaker. By studying historical data, we can evaluate which arguments are effective to which others according to the belief of the speaker.

With the absence of knowledge, it is usually very hard for deceivers to infer the same stories as truth tellers. This means no matter how convincing a deceiver's story sounds, his selection of arguments is different from a truth teller's. For the purpose of persuasion, it is reasonable to raise the most critical and relevant arguments, but deceivers fail because they do not have a complete view of the entire story, and thus they may leave out some subtle but critical arguments. As a result, some arguments may sound overly certain while others may sound unconvincing. Deceivers heavily manipulate some arguments probably because they believe that these arguments can convince the listeners of the conclusions, and consequently these arguments, after manipulations, sound more reasonable in support of the conclusion. Nevertheless there are also exceptional cases in which deceivers cannot avoid overly manipulating arguments that are usually ignored by truth tellers. We call them Type I incredibility: incredibility due to over-manipulation. The arguments that are not convincing usually can be found in the arguments that were slightly manipulated or ignored by the deceiver because deceivers do not know that they are important supports to the conclusion but truth tellers never neglect these details. This is called Type II incredibility: incredibility due to ignorance. Type I and Type II incredibilities are two examples of unconvincing arguments (According to DePaulo et. al (2003), liars tell less compelling tales than truth tellers), which can be quantitatively measured in the semantic arguments. We summarize the incredibility types in Table 6. In the law case example, Type I incredibility does not occur, but Type II incredibility appears in the argument *B_is_new_to_party* and *B_refuse_to_pay*. The deceiver ignored these arguments, which results in the incredibility of the story. On the other hand,

misinformed stories do not show this correspondence between inconsistency and untruthfulness.

Table 6 Incredibility types of manipulated arguments ($S_{original}$ refers to the value of the original state of argument, S_{truth} refers to the value of the true state of argument, and $S_{manipulation}$ refers to the value of the manipulated state of argument)

Type I	Incredibility due to over-manipulation	$S_{original} < S_{truth} < S_{manipulation}$ or $S_{original} > S_{truth} > S_{manipulation}$
Type II	Incredibility due to ignorance	$S_{original} < S_{truth} \cap S_{manipulation} < S_{truth}$ or $S_{original} > S_{truth} \cap S_{manipulation} > S_{truth}$

These patterns indicate that the uniqueness of the deceivers' reasoning process lies with not only the exhibition of inconsistency and untruthfulness, but also with a specific distribution of inconsistency and untruthfulness within the knowledge base. Propagated manipulations, functionality and dependence between inconsistency and untruthfulness are the guiding concepts of our deception detection method, which forms a general model independent of the domain knowledge. Among all three patterns, only functionality of arguments has been studied in earlier research (DePaulo et al., 2003). Often, the functionality is measured subjectively by human raters, but here we can derive and explain all the patterns theoretically and measure them computationally based on the cognitive model of argumentation.

We compute the scores of the patterns that indicate their strengths in a story. The first pattern, propagation of manipulation, depends on the relationship or dependence between arguments. The dependence of two arguments can be measured by the concept of mutual information (Li, 1990), which is a quantity that measures the mutual dependence of two variables. The dependence between two variables with discrete states can be obtained using the following equation (Li, 1990):

$$I(A,B) = \sum_{i,j} p(A_i B_j) log \frac{p(A_i B_j)}{p(A_i)p(B_j)}$$

where $I(A,B)$ represents the dependence between argument *A* and *B*, *i* refers to the i^{th} state of *A*, and *j* refers to the j^{th} state of *B*. $I(A,B) = 0$ if and only if *A* and *B* are independent. This is easy to see in one direction: if *A* and *B* are independent, then *P(AB) = P(A)P(B)*. The higher $I(A,B)$ is, the more dependent *A* and *B* are, and the more likely that manipulations propagate from one to the other. The observation that manipulations propagate to dependent arguments indicates that inconsistency of an argument is contributed to by its influential neighbors. Therefore, to measure the propagation, we first estimate the inconsistency of an argument contributed by its dependent arguments, and then calculate how strongly the exact inconsistency correlates with the contributed inconsistency. The contributed inconsistency of an argument can be calculated as the sum of the inconsistencies of all other arguments weighted by the argument's dependences on them, which is:

$$contribution(A_j) = \sum_{i<N \& i \neq j} inconsistency(A_i) \times I(A_j, A_i)$$

where $contribution(A_j)$ denotes the contributed inconsistency of argument A_j, and N denotes the total number of arguments.

The score of the pattern is calculated as:

$$r_{A,manipulation} = \frac{cov(contribution(A), inconsistency(A))}{\sigma_{contribution(A)} \sigma_{inconsistency(A)}}$$

where $r_{A,manipulation}$ denotes the correlation between the exact inconsistencies and the contributed inconsistencies of the nodes in the set *A*, and $inconsistency(A_i)$ denotes the inconsistency of argument A_i.

Being functional to the conclusion indicates that an argument is close to the expected value based on its relationship with the conclusion and the assumptions. Thus, functionality can be calculated by predicting the value of an argument using the value of the conclusion and the values of the evidence, which is:

$$\hat{p}^t_{functionality}(A_i) = \overline{p(A_i)} + \frac{\sum_{j \in E \cup c}(p^t(A_j) - \overline{p(A_j)})r_{ij}}{\sum_j |r_{ij}|}$$

where $\hat{p}^t_{functionality}(A_i)$ denotes the probability of the argument A_i predicted by the conclusion and the evidence in repeat t, $p^t(A_j)$ denotes the probability of the argument A_j in repeat t, $\overline{p(A_j)}$ denotes the average of the probabilities of A_j over all repeats, and r_{ij} denotes the Pearson Correlation between A_i and A_j. E is the set of evidence and c is the conclusion. For the purposes of presentation, we will name the argument predicted by the goal and the evidence as the functional argument and the argument predicted by the correlation network as the expected argument. An argument is functional to the conclusion and the assumptions if the actual argument leans towards the functional argument instead of towards the expected argument, or computationally, the probability of the actual argument and the probability of the functional argument are on the same side by reference to the probability of the expected argument. We calculate the functionality of a story as the ratio of arguments that are functional, which is:

$$functionality(A)$$

$$= \frac{\sum_{i \in Inconsistent(A)} 1\{(\hat{p}^t_{prediction}(A_i) - p^t(A_i))(\hat{p}^t_{prediction}(A_i) - \hat{p}^t_{functionality}(A_i)) > 0\}}{|Inconsistent(A)|}$$

where $functionality(A)$ denotes the functionality of the story A, $p^t(A_i)$ denotes the probability of argument A_i in repeat t, $\hat{p}^t_{functionality}(A_i)$ denotes the probability of the

functional argument of A_i , $\hat{p}^t_{prediction}(A_i)$ denotes the probability of the expected argument of A_i, and $Inconsistent(A)$ denotes the set of inconsistent nodes in the story A.

The negative association between inconsistency and untruthfulness indicates that if an argument is inconsistent it is unlikely to be untruthful. To measure this pattern we compute the fraction of inconsistent arguments that are not untruthful as

$$correspondence(A) = \frac{|Inconsistent(A) \cap \overline{Untruthful(A)}|}{|Inconsistent(A)|}$$

where $correspondence(A)$ represents the dependence between the inconsistencies and the untruthfulness in the story A, and $\overline{Untruthfulness(A)}$ denotes the arguments that are not identified as untruthful by the consensus network. A higher proportion indicates a stronger dependence.

3.2.3 Module III: Classification

Based on the scores obtained from the module for reasoning pattern identification, we form a feature vector with three dimensions to represent the reasoning patterns in a story, and label each story with deception or truth. As in conventional methods of detection, supervised classifications are used to learn the model of classification and map unseen data to the categories of deception and truth. The predictive performance of the classifier is evaluated using 10-fold cross-validation. In our framework, we select logistic regression for the classification due to its simplicity instead of the commonly used classifiers of Naïve Bayes and Support Vector Machine (SVM). Logistic regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. It generates a probability score to

indicate the likelihood of the category of a test case. We use logistic regression to predict how likely a story is deceptive depending on the feature vector of the reasoning patterns.

3.2.4 Summary

We proposed a model to detect deception using three steps. In the first step, inconsistency is identified by the deviation from the speaker and untruthfulness is identified by the deviation from the truth tellers. These deviations bring unexpected observations to the attention of the detector, and their distributions provide further explanations about the speaker's reasoning. In particular, if the manipulation propagates to related arguments according to dependence, the manipulated arguments are functional to the claimed conclusion; and, if there is little overlap between the inconsistent arguments and the untruthful arguments, then the story is likely to be deceptive. Based on the strengths of the patterns, a trained model of classification classifies the story into the category of deception or truth.

We realize that our model shares similar processes with Johnson et al.'s (2001) model. The first process in Johnson et al.'s model catches the discrepancies as we did in our model. The discrepancies are identified through domain-specific knowledge, which is financial knowledge in Johnson et al.'s case, whereas in our model, expectations of domain knowledge are represented by the context knowledge obtained from correlated agents other than supplied by domain experts. In addition to the deviation from the group norm, we require the stories to be consistent with the speaker's historical data because the deviation from one's own belief is a necessary component of deception. In their second

and third processes, hypotheses are proposed and evaluated to explain the discrepancies. The hypotheses are generated according to the taxonomy of deception tactics and some domain-specific goals. We explain the discrepancies by evaluating the reasoning patterns involving the discrepancies. Since the patterns are derived from the types of reasoning instead of from the deception tactics in a specific domain, the patterns can be generally applied to any topic. The last process that combines all accepted hypotheses to produce a final outcome can be mapped to the classification module in our model. Driven by the intent of deception, we are able to take the modeling of the detection processes one step further by generalizing methods and removing assumptions. Since the manipulation of arguments in a communication is so flexible that there are innumerable ways to deceive, we look for the universal patterns of deception and verify indicators through domain-independent criteria. On the contrary, Johnson et al. customize the expectations of observations and map indicators to a limited number of deception tactics based on domain knowledge. Our model also performs detection on data with a more flexible form than the accountant reports in Johnson et al.'s case. Overall, our model of detection generalizes Johnson et al.'s model. It is completely domain independent and does not assume the form of data as long as it provides text-based content with sufficient levels of reasoning.

3.3 Outcome of Detection

The final output of the model includes a quantified result indicating the likelihood of deception in a story. Quantified results are not only more informative but also more

reasonable because telling the whole truth or a whole lie is rarely a desirable action in the real world. A qualified (binary) result can be obtained by thresholding.

The contribution of our model lies with its detection power and its analysis capability, a combination of which we have rarely found in any existing work. This analysis capability includes deception editing, deception explanation and hypothesis evaluation. Deception editing restores the honest arguments of a deceiver using GroupLens prediction method. Deception explanation explains why a story is deceptive in terms of (i) how the deception was formed and (ii) how the deception was detected. To explain how the deception was formed, a hierarchy of the manipulated arguments in a deceptive story can be constructed to illustrate the flow of manipulations. Together with a comparison between the deceptive arguments and their honest versions, the type of manipulation can be revealed. To explain how the deception was detected, we analyzed the strength of the reasoning patterns discovered from a story. In particular, we can evaluate the score and the activation of each pattern. These analyses can provide human detectors with valuable evidence of deception in that they can further review and/or confirm the decisions. In our model, the reasoning patterns are not limited to the three we proposed. Any quantifiable hypotheses regarding human reasoning can be incorporated into the model and serve as features to classify deception. As such, hypotheses on the patterns of deceptive reasoning can be evaluated in a computational way. Embodying this function, the capability of hypothesis evaluation can also assist the psychological and cognitive studies of deception.

Chapter 4 Detecting Real World Deceptions

4.1 Overview

We have argued and verified through experiments that the major discrepancies between deception and truth lie with the inconsistency of a speaker's behavior and his untruthfulness compared with truth tellers. The discrepancies are measurable through computational models. We have also described and credited three reasoning patterns of a deceiver that non-deceptive communications do not exhibit: manipulations propagate to related arguments according to dependence, manipulated arguments are functional to the claimed conclusion, and inconsistency and untruthfulness are negatively associated. Our earlier verification of the hypotheses was based on an artificial scenario. Here, we will attempt to evaluate the hypotheses again using real world cases.

When applying the model to real life scenarios, we will face practical problems that have not been faced when just using word-level cues. Firstly, it is a challenging task to elicit the semantic arguments from natural-language communications. There are several problems we need to address during the process, such as the retrieval of the semantics from a natural-language story. We may rely on natural language processing (NLP) tools to tackle the problem directly, but tools that can process subjective information as sensibly and precisely as a human would are not available. Secondly, historic data from multiple agents are required in our framework in order to measure one's deviation from the self. Although this kind of data exists in the real world, it is seldom recorded especially for the purpose of deception detection because people avoid admitting

deceptive acts, and detection by others lacks ground truth. Unfortunately, it is also unusual for surveys to record historical information because individual differences have rarely been considered in past studies. (There are no longitudinal studies that we are aware of.) A third obstacle is that real world data is noisy. Real world deceivers may not strictly follow the reasoning process we proposed. Psychological status and communication environment mediate the process of argumentation although consciously making changes to the reasoning process is very difficult. Most of the problems are within the realm of NLP and intent modeling. Although NLP and intent modeling are not the focus of this thesis, we will provide a comprehensive analysis of possible methods.

The outline of our experiments and evaluations are presented in Fig. 6 and Fig. 7. To apply our detection method on existing real world datasets, we need to synthesize data in a reasonable way. To this end, we first build the cognitive models of the speakers according to the stories in the datasets, and then synthesize stories by conducting inference on the models. The evaluation of the cognitive model includes a parametric study that analyzes the sensitivity of the model to different assumptions during its construction. The evaluation of the detection model is accomplished through evaluating the inconsistency detection, evaluating the detection of deception, and discussing analysis capability. For inconsistency detection, we used both simulation data and real data to measure the performance. In deception detection we first evaluate detection performance on classifying truth and deception with real data, and then synthesize misinformed agents to evaluate its capability to distinguish deception from misinformation. To study the behavior of the detection model we also perform parametric studies to evaluate how

inconsistency detection and deception detection react to different types of datasets. In addition to the detection capability, we also present the analysis capability of the model which enables the restoration of truth and the explanation of deception.

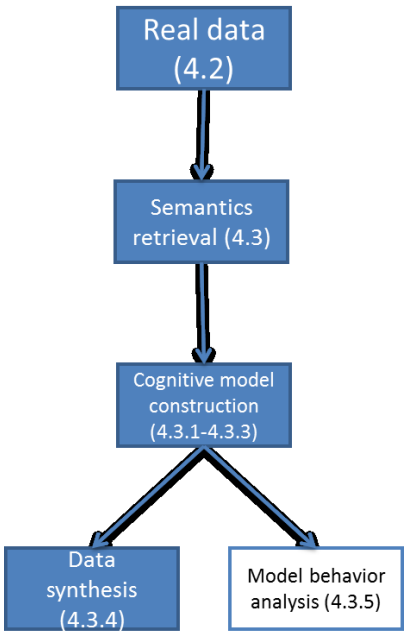


Figure 6 Outline of the experiments in the construction of cognitive models

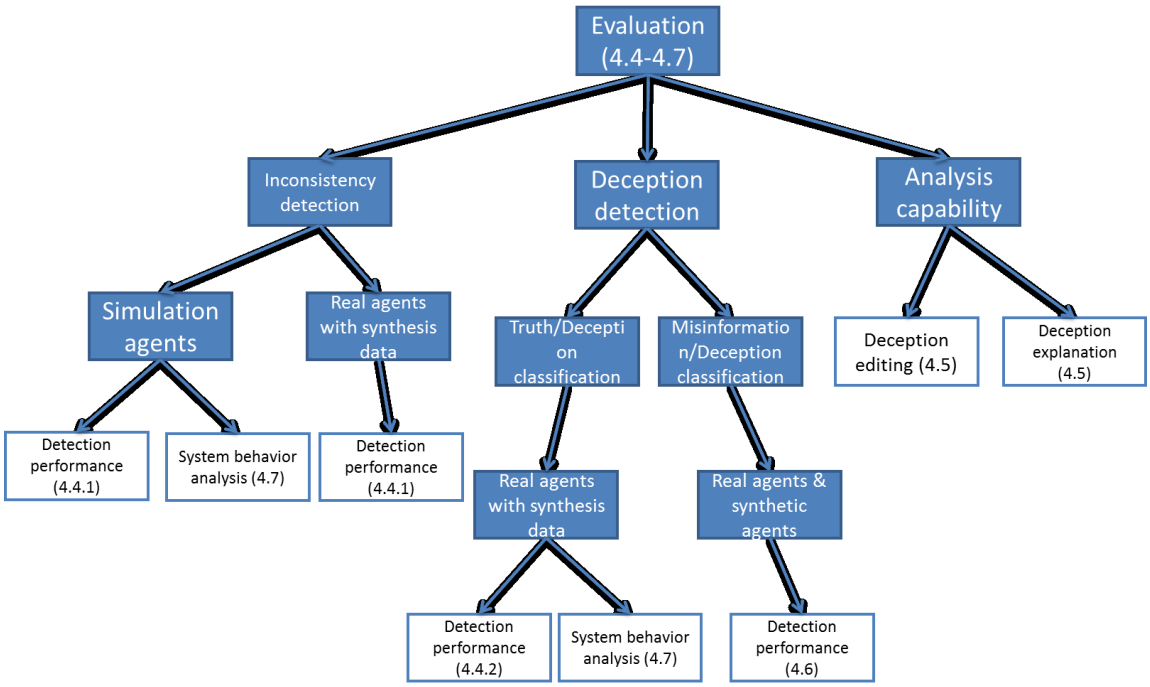


Figure 7 Outline of the experiments in deception detection

4.2 Deception Datasets

According to Gokhman et. al (2012), deception datasets commonly used in the field of deception detection can be categorized into sanctioned deception and unsanctioned deception. In sanctioned deception the experimenter supplies instructions to individuals to lie or to tell the truth. In unsanctioned deception, the participant lies of his or her own accord. The majority of studies on deception detection employ sanctioned datasets in which experimenters recruit participants for a lab survey and randomly assign them to a lie or truth condition. Sanctioned datasets are easy to control, and thus are able to provide clean data with ground truth. However, they are found to be not as realistic as unsanctioned datasets. In reality, people are motivated to lie instead of instructed to lie. The benefits of success and the risks of failure of deception strongly influence the performance of deceivers. Unsanctioned datasets address this problem by asking the participants to record their daily lies, but the participants may not always completely recall the lies they made and sometimes unconsciously deny the lies. Another type of unsanctioned deception involves incentivizing the participants to cheat in an environment and to lie about the cheating later. This type of data collection is very labor intensive and usually yields very small number of lying cases. In this thesis, we use lab surveys as our datasets since lab surveys provide clean data in which the participants are not influenced by the different testing environments, sufficient information about the survey can be provided for analysis, and ground truth are available for the purpose of evaluating performance.

Among all lab surveys that are available for research purposes, Mihalcea et. al's surveys of opinions on controversial topics (2009) and Ott et. al's hotel reviews (2011) are most popularly used by existing researchers as benchmarks. We choose their datasets for the purpose of facilitating the comparison of our methods with existing methods. In Mihalcea et. al's surveys, 100 participants were asked to imagine that they were taking part in a debate. To prepare for their speech, each of them needed to provide a story of at least 4 or 5 sentences to illustrate his true opinion on a topic. Next, they were also asked to provide stories to illustrate false opinions on the topic, that is, to lie about their opinions. For example, on the topic "abortion", participants may argue about why abortion is good if they support abortion, and also argue about why it is bad as if they did not support it. Among the three topics they surveyed ("abortion", "death penalty" and "my best friend"), we select the data under the topic "abortion" because there are sufficient numbers of both positive and negative opinions on it in both true and deceptive stories, which provides us with some information of how arguments from different sides influence the conclusion, or in other words, enable us to learn the reasoning behind the comments, and at the same time prevent us from being biased to one side of the story. The hotel reviews from Ott et al. are composed of 800 true reviews collected from TripAdvisor for 20 hotels in the Chicago area and 400 deceptive reviews gathered using Amazon Mechanical Turk (AMT) for the same 20 hotels. To make sure that the hotel reviews are comparable to the opinion surveys, we randomly selected 5 hotels to form a corpus of 100 deceptive reviews and 200 true reviews. Samples of the abortion data and the hotel reviews can be found in Appendix E. In both surveys, the participants were asked or assumed to provide all true arguments in the true stories, but were not required to provide all lies in the

deceptive stories. This means that we only have ground truth for global lies, that are lies in the conclusion of a story, but not for local lies, that are lies in their individual arguments. The major difference between the opinion surveys and the hotel reviews are: opinion surveys provide opinion-based stories that are subjective beliefs on a topic while hotel reviews provide fact-based stories which are concluded based on facts and experience; and in opinion surveys each participant provides a true story and a deceptive story while in hotel reviews true stories and deceptive stories are provided by different speakers. The first difference also refers to the difference between opinion-based stories and fact-based stories. Opinion-based arguments are subjective beliefs about ethics or feelings of the speaker, which rarely changes according to observations due to the cognitive dissonance (Festinger, 1962) aroused by the conflict between facts and opinions. Depending on the topic, they may or may not contain objective facts. Opinion-based stories, especially those on controversial topics, can be very different from person to person such as the diverse views on abortion. It is hard to detect deception in opinion-based arguments because speakers usually can argue on both sides without the support of facts. As a result, deceivers are not necessarily less compelling than truth tellers. On the other hand, in fact-based stories speakers usually hold similar attitudes towards the same observations such as the sentiment on a large hotel room, the attitude on a low price and the attitude on a heavy laptop. However, speakers' attitudes and conclusions are sensitive to observed facts if they do not have a strong bias towards one side of the story. Thus, we can expect to see an improvement in the performance of detection in fact-based stories compared with opinion-based stories. Mihalcea et al.'s work demonstrates that their detection performance on the dataset involving both facts and opinions (my best friend) is

significantly better than the performance on the purely opinion-based stories (abortion and death penalty). Testing on both opinion-based and fact-based stories gives us a good spectrum on the performance of our model.

One major limitation of existing deception datasets is that historic data and/or multiple stories from each participant are not available, and it is this information which deception detection should rely on to establish the original beliefs of each participant unless the participants' opinions can be predicted by human experts or other mechanisms. To cope with this problem, we model the entire cognitive process of argumentation and synthesize historic data for each speaker in both datasets. A cognitive model of a speaker is called an agent in our experiment.

Another limitation in the hotel reviews is that the true stories and the deceptive stories are provided by different people. We require paired stories of truth and deception because it is necessary to know one's history in order to measure his inconsistency. When testing the opinion surveys, we can directly classify deceptive stories from true stories, but when testing using the hotel reviews we cannot use the stories directly. Instead, we use the true hotel reviews to generate individual agents and use the agents to synthesize true stories and deceptive stories. The synthetic true stories should be similar to the original true reviews because the agents were learned based on the original reviews. To ensure that the detection on the hotel reviews is comparable to that on the opinion surveys we picked 100 agents whose stories contain the highest levels of reasoning as determined by computing the number of edges between the explicit arguments in their stories. Deception

with higher level of reasoning is more common among experienced deceivers, and thus is more difficult to detect. We will discuss the level of reasoning in detail in Chapter 4.7.2.2.

4.3 Data Synthesis and Cognitive Process Modeling

In real time communication environments where the history of each speaker can be recorded only techniques that retrieve semantic arguments from communications are enough for data gathering, but in other areas where history data is inaccessible, intent modeling techniques that can synthesize stories are necessary. Unfortunately, as our literature review revealed, none of the existing datasets contains multiple true stories from a participant because real life deception data over a long period of time are very rare and any survey data would require massive human effort. Usually each participant provides one true story and/or one deceptive story. Synthesizing stories is a challenging task because we need to make sure these stories are consistent with the stories in the dataset, or in other words, the synthesized stories are likely to be produced by each individual participant of the survey, and at the same time sound realistic and reasonable. To this end we model the cognitive process of the participants by learning their reasoning processes and generating artificial knowledge bases, (BNs in our case), accordingly assuming that the datasets are the presentations of their reasoning results, followed by the inference of possible stories and the selection of arguments through the use of their cognitive models. Learning of the reasoning process or even retrieval of the semantic meaning from natural language is a major challenge for NLP. The difficulties in our work particularly include the identification of relevant arguments, the extraction of polarities on arguments, and the construction of the relationships between arguments.

As discussed in Chapter 2.3, relationships between arguments are encoded by graphical representations of BNs in the cognitive model. A node in a BN represents an argument, which is instantiated by two states referring to the positive (believe) and negative (not believe) polarities of the argument. To retrieve the semantic arguments from the stories we use rule-based keyword mapping combined with some manual effort. Specifically, we pick out the most frequent *K* arguments from all true stories by manually picking out the arguments containing the most frequent *N* words. As an example, the arguments under the abortion dataset are listed in Table 7. We use a stemmed word to represent the key to each argument, and the argument "abort" is the conclusive argument for stories in the abortion data. To retrieve the polarities of the arguments from each speaker's story, we need to first identify the arguments and then extract their sentiments. Both processes are accomplished through mapping phrases with argument related key words (e.g. "woman", "right", "bodi" are key words related with the argument "right") and sentiment related key words (e.g. "not", "couldn't" are key words related to the negative polarity). An agent's polarity on an argument can be positive, negative or missing. We use a vector of real numbers to represent the arguments in a story with 1 denoting the positive polarity (e.g. fetus is human), 0 denoting the negative polarity (e.g. fetus is not human) and a special number (usually 0.5 or -1) denoting a missing argument as in:

$$Story_i = \{x_1\ x_2\ x_3\ \dots x_N\}$$

where $Story_i$ denotes the story of the i^{th} agent, N is the number of arguments, x_j represents the polarity of the j^{th} argument, and $x_j \in \mathcal{R}$. An entire dataset can be represented as

$$Dataset_{topic} = \begin{Bmatrix} Story_1 \\ Story_2 \\ \dots \\ Story_M \end{Bmatrix}$$

where $Dataset_{topic}$ denotes the dataset of topic *topic*, and *M* is the total number of stories. The true dataset and the false dataset can be respectively represented as

$$Dataset^{True}_{topic} = \begin{Bmatrix} Story_1 \\ Story_2 \\ \dots \\ Story_{M/2} \end{Bmatrix} \qquad Dataset^{False}_{topic} = \begin{Bmatrix} Story_{\frac{M}{2}+1} \\ Story_{\frac{M}{2}+2} \\ \dots \\ Story_M \end{Bmatrix}$$

where $Dataset^{True}_{topic}$ denotes the true dataset of topic *topic*, where $Dataset^{False}_{topic}$ denotes the false dataset, $Story_i$ represents the true story of the i^{th} agent, $Story_{\frac{M}{2}+i}$ represents the false story of the i^{th} agent, and *M/2* is the total number of true stories and false stories. After the generation of the polarities, we carefully go through all the stories to confirm the mappings. We have detailed the process of data elicitation in Appendix C.

Table 7 Arguments in the abortion dataset

Abort (conclusive argument)	I support Abortion.
Right	Women have the right to do whatever they want with their bodies.
Govern	Government should interfere with people's decision on abortion.
Care	Unwanted children are put into unregulated care systems.
World	Unwanted children should be brought up in the world.
Life	Unwanted children's lives are miserable.
Murder	Abortion is murder.
Health	Some pregnant women have health problems.
Option	Abortion is an option.
Time	The time to allow abortion should be fixed.
Early	Abortion should only be allowed at early time.
Population	Abortion can help birth control.
Adopt	Adoption is an option.
Rape	Some pregnant women were raped.
Carry	Women should be forced to carry babies.
Child	Children have the right to life.
Couple	There are families and couples who want to adopt babies.
Educate	Education should be provided to prevent unwanted pregnancy.
Mistake	People use abortion to correct their mistakes.
Teenager	Some teenagers get pregnant.
Inconvenience	Pregnancy is inconvenient.
Responsible	People should take responsibility.
Sex	People are forced to have sex.
Birth	Some pregnancies have birth defects.
Human	Unborn children are human.
Concept	Life starts from conception.
God	Religion plays an important role in the decision.
Circumstance	There are circumstances when people need an abortion.
Want	People want abortion.
Legal	Abortion is legal.

4.3.1 BN Learning

The first task in the agent modeling is to build a BN for each speaker based on the stories in the dataset. The agents are represented by BNs with the same structure, whose parameters are respectively learned from the agents' individual stories. The task of BN learning is achieved through two steps: building the graphical structure of the BN and

filling in the numerical parameters of the BN, which are the conditional probability tables (CPTs). This problem has been studied in depth over the last decade, and consequently there is a considerable number of learning algorithms. Existing methods to build BN structures can be grouped into two categories: methods based on score functions and search, and methods based on independence tests. Score and search based methods are more frequently used in learning BN structures. The algorithms attempt to find a structure that maximizes the fitness between the graph and the data measured by a scoring function. Search algorithms are combined with the scoring function to explore structures with high scores within a space of feasible solutions. Brute-force search is usually unaffordable. Therefore heuristic search algorithms were proposed to efficiently reduce the search space. The advantage of score and search based methods is that they usually can find the global optimal solution, but the computational complexity can be extremely high. Compared with the quantitative approach of scoring and searching, methods based on independence tests build BNs through qualitative studies of prior knowledge in the domain. They discover causal relationships through analyzing the dependence between random variables (r.v.s) and form graphs that encode assertions of conditional independence. The algorithms are able to build the structures more efficiently by following the intuition that the structure of a BN corresponds to dependencies implied by the distribution subject to the observational data. The algorithm we use is an independence test based algorithm called PC algorithm (Sprites, 2000). PC algorithm is widely used to generate BNs from correlations and partial-correlations of observations. The basic idea of PC algorithm is to test the conditional independence using d-separation. If a path connecting a node u and a node v is d-separated by a set of nodes Z, u and v are

said to be conditionally independent. The d-separation can be tested using the partial correlation between *u* and *v* given *Z*. After the nodes are connected based on d-separation, heuristics that avoid cycles are applied to direct arcs. The advantage of using PC algorithm is that it conforms to the intuition that causally related arguments are always correlated and that it saves computational and memory costs of searching a huge space of all feasible structures while still being able to distinguish causality from correlation. Assuming all agents share the same structure, we use $Dataset^{True}_{topic}$ to generate a BN in which each node represents an argument. The true stories from 100 agents provide us with a good measure of the correlations and partial correlations between arguments, which are further used in the PC algorithm to derive causal relationships. We have detailed the process of learning BN structure in Appendix C. During the process of BN learning, we varied the threshold of correlation under which the edge between two arguments should not be preserved. A high correlation threshold means that two arguments are allowed to be causally related only when they are strongly correlated. We select the BN learned from the threshold that can minimize the number of arcs without disconnecting the graph. Figure 8 depicts the graph that we built from the abortion dataset.

The objective of BN learning, in general, is to construct the BN that best fits the learning data, which is accomplished by finding the parameter that can maximize the likelihood of data given a structure. In our experiment the parameter of each agent's BN is individualized according to his personal beliefs. Obviously deceptive stories cannot be included in the learning data because they do not represent the true beliefs of the speakers.

This leaves us with one piece of data, (the true story), for each BN, which means we need to learn parameters from extremely sparse data. Also arguments are missing in each story, so our data is incomplete. Theoretically, there can be an infinite number of BNs fitting one piece of data. However, we require additionally that the BN realistically represents a human speaker, which means that it should (i) most likely generate the speaker's story, (ii) represent the corresponding speaker better than any other BN, and (iii) consider other possibilities of the speaker's attitude. We are able to fulfill the first requirement by maximizing the likelihood of the speaker's data. The second and third requirements are actual conflicting. The second requirement can be fulfilled by maximizing the diversity of the agents and minimizing their uncertainties, which implies that the agents are extremely biased by their original stories and are reluctant to change with different evidence. In this case, the third requirement that speakers possess alternative attitudes cannot be fulfilled, therefore the learning algorithm is required to balance the bias of the agents by favoring the agent's story while sharing knowledge with others. Why do we want a model that favors one story but also considers other possibilities? A model represents the reasoning process of a speaker. If a model is very biased to the speaker's own story it indicates that the speaker's reasoning is deterministic, and consequently that the speaker will strongly insists on the same arguments no matter what evidence is observed, which is a very unrealistic scenario. On the other hand, an extremely certain model will artificially ease the detection of deception. Since the model always biases to the same arguments, the training data becomes very similar and the prediction turns out to be precise and certain. As a result any deviation from the true stories is regarded as inconsistent. By considering more possibilities of arguments the agent becomes hard to

predict which then maximizes the difficulty of deception detection. Modeling speakers in this way we can have an estimate of the lower bound of our detection performance. This requirement in modeling suggests two considerations. Firstly, people reason with uncertainty. Although the story is the speaker's best selection of arguments, he may have sub-optimal selections, and his preference on the sub-optimal selections may not deviate significantly from the best one. An ideal model should appropriately encode uncertainty. In addition, a speaker's model should incorporate context knowledge of the domain which is not explicitly mentioned in the speaker's story since the context knowledge may shed light on the other possible selections of arguments. The considerations elicit two more constraints in the learning of parameters: maximize the uncertainty of the agent, and use data from all agents as a reference of context knowledge. Overall, we want to build personalized models with uncertainty based on a common ground of knowledge instead of independent models that can best fit their individual stories. Note that we do not require the models to avoid fitting the deceptive stories because it will artificially increase the detection rate.

We determined that an approach based on the principle of maximum entropy can satisfy our constraints. The approach handles missing data in the following way: "when we make inferences on incomplete information, we should draw them from the probability distribution that has the maximum entropy permitted by the information which we do have", (Moens, 2006). According to this idea the joint probability of the "actual argument" and the "presented argument" can be estimated using the maximum entropy principle, subject to the constraint that the marginal for the "presented argument" agree with the

maximum likelihood estimates (MLE) of the learning data. Under the assumption that data are missing at random, the Maximum-entropy algorithm is equivalent to a localized EM algorithm, (Cowell, 1999). Entropy-based algorithms specify two constraints: maximizing the likelihood of learning data, and maximizing or minimizing the entropy of the model. By maximizing entropy we interpret the reason for missing data as because the speaker is uncertain about it. Minimizing entropy can be interpreted in two ways: the data is ignored because the speaker is very certain about it (or it is common sense), or the data is not mentioned because the speaker does not believe in it.

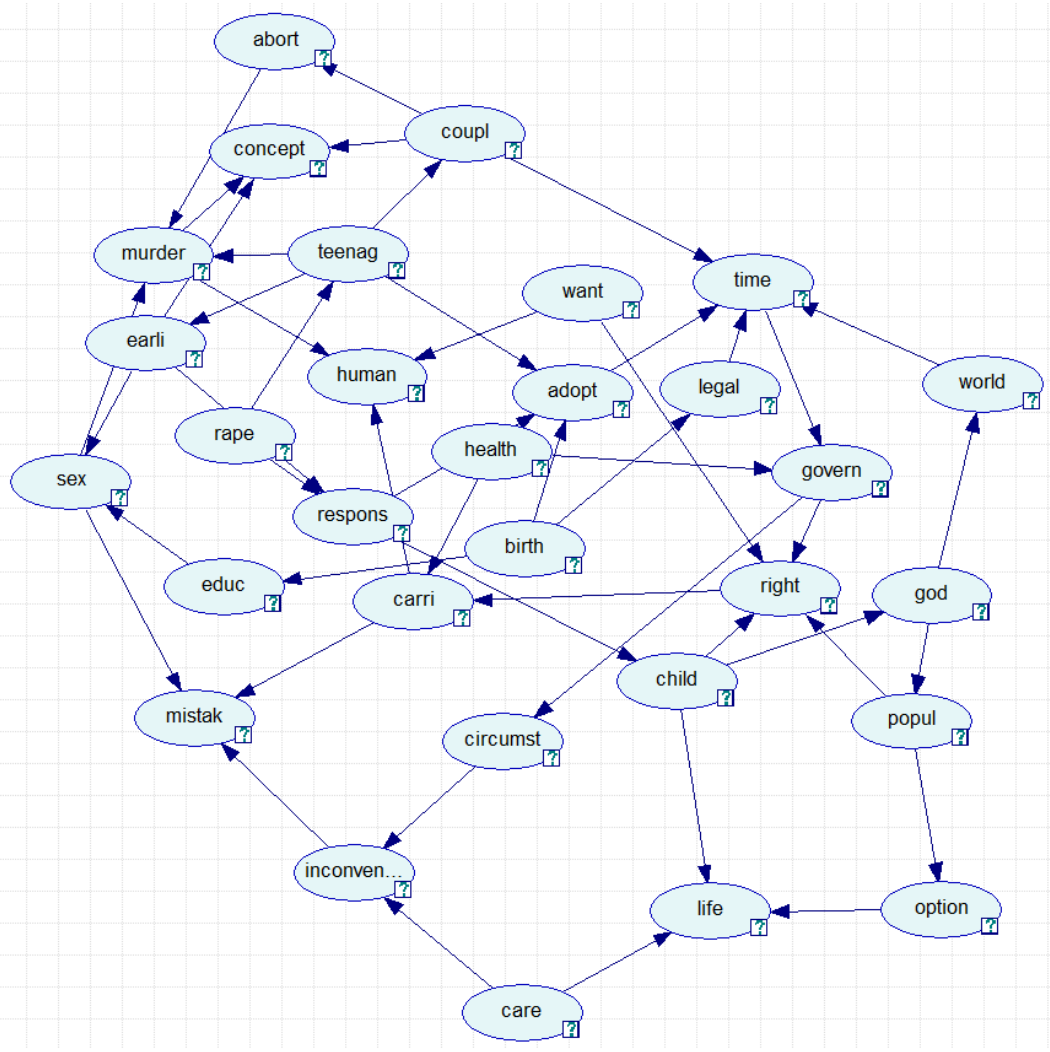


Figure 8 Agent generated from one hundred true stories from the abortion data

When learning the parameter for the i^{th} agent we try to enrich the context knowledge of his model by using the entire $Dataset^{True}{}_{topic}$ as learning data, but with a lower weight relative to $Story_i$. We want to find the minimum weight of his story that can be used to build a BN generating his own story as the MPE. The intuition behind this method is as follows: Assuming that people's knowledge on daily topics is learned from similar observations, we can expect them to share similar context knowledge, which is also called common sense. People also form their individual opinions because a particular observation(s) was enhanced more frequently than others, and the most frequent observation(s) becomes the most likely story that one believes.

4.3.2 Argument Selection

A BN learned based on semantic arguments using the Maximum-Entropy based algorithm represents a rational person who infers the given data as his most likely story without losing generality to the common sense of the majority of people. However, when his inferred story is communicated, it is usually presented as binary polarities, which are "believe" (positive), "not believe" (negative) or "unknown", instead of by probabilities. Positive and negative polarities can be easily identified by favoring the state with a higher probability, but not all explicated arguments are presented. Although speakers may have opinions on all arguments, they may not mention certain arguments for some reasons. As we have presented in Chapter 2.3.3, cognitive scientists (Carenini and Moore, 2006) suggested that an argument is worth mentioning if it is significantly more compelling to a target argument compared with other arguments. We use the method proposed in Chapter 2.3.3 to measure the compellingness of arguments. An argument is notably-compelling,

(worth-mentioning), if its compellingness exceeds all other arguments' by k std. deviations. When determining the notably-compelling arguments given k we build a hierarchy of mutual impact starting from the conclusion. Specifically, we insert the conclusion in the lowest level of the hierarchy. For each argument on the current level we calculate the compellingness of all other arguments to it and insert the notably-compelling arguments in the next level if it is not in the hierarchy yet. Then, we move to the next level and continue the process until no new argument is added. Fig. 9 depicts a hierarchy of mutual impact built in this manner. This is a reasonable way to elicit compelling arguments because the purpose of the arguments is to support the conclusion. An argument is worth mentioning only if it can support the conclusion or support other arguments supporting the conclusion.

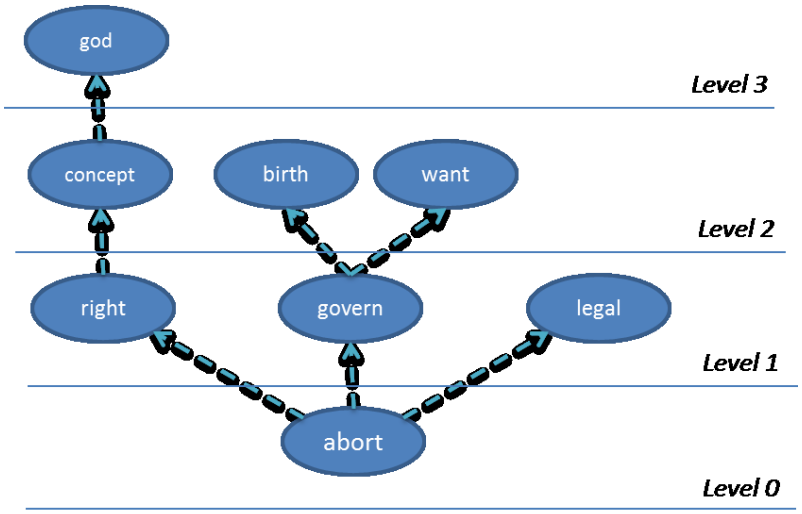


Figure 9 A mutual impact hierarchy to determine notably-compelling arguments

The value k in determining notably-compellingness can be used to represent a speaker's personal tendency to believe in an argument. We measure k for each agent by

maximizing the likelihood of representing explicit arguments as notably-compelling and implicit arguments as not notably-compelling.

When reasoning through a BN, observations and assumptions can be set as evidence. When evidence is not explicitly pointed out in a dataset we need to select evidence manually. Selection of evidence is not as straightforward as it seems to be. Evidence means objective truth, but even in fact-based stories like hotel reviews people's perceptions of facts can be different. For example, some customers may regard two bathrooms as a plus but others may feel it redundant. The size of a bed and the rate of a room can be judged differently according to people's preference, therefore, we identify an argument as evidence only when all speakers in the same environment agree on it. Specifically, for a topic with the same setting (e.g. the hotel reviews for the same hotel), if all speakers who explicitly mention an argument agree on the argument, it is identified as evidence. An example of evidence is "The hotel charges extra fees on facilities". Some other arguments are set as evidence manually because they are premises or events. An example is "We stayed more than 3 nights".

4.3.3 *Validation of Cognitive Model*

A cognitive model has a representational power if its inferred true story is similar to the original speaker's true story and its inferred deceptive story is similar to the speaker's deceptive story. To infer a true story the corresponding agent's BN is reasoned given the evidence of the true story. To infer a deceptive story, both the original evidence and the pre-assumed conclusion are set as evidence, then the inferred arguments are classified as

explicit (mentioned), or implicit (missing), according to the significance of their compellingness. To evaluate the representational performance of the cognitive models all stories are inferred and the similarities between the inferred stories and the original stories are measured. The performance is calculated as the fraction of the actual arguments that have the same polarities as the inferred arguments. Firstly, we validate that our deceptive reasoning is far better than random guessing. If a person randomly picks polarities of arguments to represent true stories, and also ensures that his conclusion is deceptive (mutually exclusive to the honest conclusion) to represent deceptive stories his performance are shown in Table 8, whereas our performance of representing the true stories in the abortion data is 0.913333 and that of representing the deceptive stories is 0.811333. Both results are significantly better than random guessing.

Table 8 Representational performance of deriving arguments by random guessing

support arguments #	2	3	4	5	10	19	29	30
Deceptive	0.5555	0.5	0.4666	0.4444	0.3939	0.3666	0.3555	0.3548
Honest	0.2222	0.25	0.2666	0.2777	0.3030	0.3166	0.3222	0.3225

In our model we do not strictly require the BNs to accurately infer the deceptive stories because we do not claim our proposed deceptive reasoning is the only strategy a deceiver takes. The deception strategy may depend on the deceiver's skill, his psychological status, and his knowledge about the listener. It may also be mediated by the environment of communication. However, we believe that the deceptive reasoning we proposed is a key process in a deceiver's cognition. Thus, we expect that the story inferred by the deceptive reasoning is better than that inferred by the honest reasoning in terms of representing the deceptive story. In particular, we want to show:

For any deceptive story *Y'*,

$$\frac{|Y' \cap arc\ max_X P(X|E, \overline{x_k})|}{|Y'|} > \frac{|Y' \cap arc\ max_X P(X|E)|}{|Y'|}$$

where $X = \{x_1, x_2, \ldots, x_n\}$ is an ordered set of states, where the order follows that of a complete set of the arguments, $P(X|E)$ refers to the joint probability of X given evidence E, and $arc\ max_X P(X|E)$ denotes the set of argument states that can maximize the joint probability of X given E, which is the MPE. Suppose the k^{th} argument is the conclusion and $\overline{x_k}$ is the state of the original conclusion, then $\overline{x_k}$ denotes the mutually exclusive state of the original conclusion, which is the state of the deceptive conclusion. This inequality means that the deceptive story is more likely to be generated by supposing the deceptive conclusion than by not supposing it. We tested this hypothesis and evaluated the representational power of the BNs using the abortion data. The test results are presented in Table 9. In the table, Hr2Hs denotes the test to represent true stories using honest reasoning, Dr2Ds denotes the test to represent deceptive stories using deceptive reasoning, and Hr2Ds denotes the test to represent deceptive stories using honest reasoning. Although the score of Dr2Ds is only slightly higher than the score of Hr2Ds, the difference between their performances is significant (<0.05). It may be because the individual difference of the models' performances is small and the performance of deceptive reasoning is in general better than that of honest reasoning. Assuming that the representational power is similar across different speakers, we doubt whether the similarity between different models' representational powers is caused by high similarities between the speakers, as a result of which personalized models may not be necessary. To test this hypothesis we evaluate whether a speaker's story can be better inferred by another speaker's model. In particular, we compare the performance of using

an agent to represent his story and the performance of using random agents to represent his story. The results are shown in Table 10, in which RHr2Hs denotes the test to represent true stories using random agents' honest reasonings, and in which RDr2Ds denotes the test to represent deceptive stories using random agents' deceptive reasonings. Table 10 demonstrates that the original model significantly outperforms the random models with regards to representing the stories, thus, we can conclude that there is a significant difference between individual cognitive processes, and that our cognitive models are able to encode the difference.

Table 9 Results of the representational performance of agents generated with the abortion data

Hr2Hs		Dr2Ds		Hr2Ds		Hr2Ds v.s. Dr2Ds
score	Std. deviation	score	Std. deviation	score	Std. deviation	Significance (p-value)
0.913333	0.109483	0.811333	0.112139	0.751	0.113949	0.0114

Table 10 Representational performance of agents generated with the abortion data compared with the performance of randomly shifted agents

RHr2Hs		Hr2Hs v.s. RHr2Hs	RDr2Ds		Dr2Ds v.s. RDr2Ds
score	Std. deviation	Significance (p-value)	score	Std. deviation	Significance (p-value)
0.782917	0.107059	0	0.80341	0.105455	3.02E-19

4.3.4 Data Synthesis

The careful modeling of the agents guarantees that we are able to synthesize reasonable and realistic stories which are likely to be believed by the speakers of the original stories. To synthesize history data for each speaker we infer the arguments based on his cognitive model with random evidence and represent the inference results as polarities according to the methods described in Chapter 2.3. The synthesis of historic data is carried out for each agent over a number of repeats. During each repeat, all agents are fed with the same

set of evidence. The synthesized data of a speaker can be represented by a matrix of polarities, with the rows denoting repeats and columns denoting arguments. This serves as the history of the speakers, which provide us information with regard to the correlation between speakers and the mutual dependence between arguments.

During deception detection each story in the dataset serves as a repeat of testing with a distinct set of evidence. For i^{th} repeat, the i^{th} story, which belongs to the i^{th} agent, serves as testing data. For any other agent, we synthesize an honest story with the i^{th} set of evidence by supposing that each agent has provided a story given the set of evidence in each repeat. The synthesis data is composed of 100×99 honest stories for the true testing data, and 100×99 honest stories for the deceptive testing data. These honest stories are used to form a consensus network for the i^{th} agent at the i^{th} repeat and to calculate the inconsistency and the untruthfulness of the i^{th} true story and deceptive story.

4.3.5 Parametric Study

When building the models we found that several parameters have a large impact on the representational power of the agents, namely the accuracy of argument retrieval, the level of uncertainty of the agents, the weight of individual's stories in learning data, the size of learning data, and the similarity between agents. We performed a parametric study to evaluate the sensitivity of a model's representational performance with respect to these parameters.

4.3.5.1 Argument Retrieval

In our current experiments the learning data, which is the semantic arguments, are generated through key word mapping along with manual effort in order to minimize the error from the retrieval of semantics. However, different raters may assign different polarities according to their respective understanding of the semantics. To minimize labor cost and avoid subjective bias in the process of argument retrieval, we can adopt NLP tools which support the automation of topic classification and sentiment retrieval. NLP methods may introduce errors to the retrieval of arguments and further degrade the representational performance of the model thus we want to test the influence from the process of argument retrieval by replacing the semi-manual method with an NLP tool that performs sentiment analysis automatically. Another purpose of this study is to evaluate the potential of automatically retrieving arguments using tools. Sentiment analysis refers to applications that extract subjective information such as attitude and emotion from text-based materials at the document, sentence or feature level. Early works in sentiment analysis, (Turney, 2002; Pang et al., 2002), were applied to the analysis of consumer reviews – whether a review expresses a positive or negative feeling. Although sentiment analysis achieves encouraging results in areas such as product reviews, it is not as accurate as human raters in understanding human opinions in general since sentiment analysis is domain-specific and its capability to retrieve subjective information from objective expressions is yet immature. We used an application called *Semantria* to perform sentiment analysis on our datasets. *Semantria* supports the extraction of sentiments based on customized concepts. Given a document it first breaks it into basic parts of speech based on POS tags. It then identifies sentiment-bearing phrases by mapping with user-supplied key words, and gives a polarity score to each sentiment-

bearing phrase according to how frequently it occurs to a set of known positive words and a set of known negative words. Since an argument can be expressed in multiple phrases the polarity scores are combined to determine the overall polarity of an argument in a text. We provide the same key words we used in argument classification to the application as seeds of arguments. Polarities retrieved by sentiment analysis may have different interpretations with ours, but the polarities are valid as long as the interpretations are consistent over all stories. The presentation performance of the BNs learned from the data generated by sentiment analysis is showed in Table 11. This (on Table 11) shows that although the performance increases in general by using automatic method, the deceptive reasoning is no better than the honest reasoning in representing the deceptive stories. We may explain this as follows, the BNs can better fit the data probably because the arguments of different agents generated by sentiment analysis are similar. Not only the learning data (true arguments) but also the deceptive arguments are similar. As a result, the learned BNs are not good representations of the individual speakers' knowledge bases because they can always fit the data regardless of the evidence. Overall the models learned from automatically retrieved data are not as representational as those learned from semi-manually retrieved data because automatically retrieved data does not encode personal difference appropriately.

Table 11 Representational performance of BNs with different semantics retrievals

Semantics retrieval	Hr2Hs		Dr2Ds		Hr2Ds		Hr2Ds v.s. Dr2Ds
	score	Std. deviation	score	Std. deviation	score	Std. deviation	Significance (p-value)
manual	0.9133	0.1094	0.8113	0.1121	0.751	0.1139	0.0114
automatic	0.9343	0.0794	0.8513	0.0904	0.8206	0.0785	0.8651

4.3.5.2 Level of Uncertainty

As we have discussed, maximizing entropy of data and minimizing entropy of data during the learning of BN handles missing data in different ways. Maximum entropy maximizes the unpredictability of an agent, while minimum entropy assumes the omission of arguments caused by certainty. In addition to entropy-based algorithms, the EM algorithm is another learning method that deals with incomplete data. EM algorithm iterates through the process of expecting missing data based on the current estimate of the parameter and the process of updating the parameter by maximizing the likelihood of learning data until the parameter converges. EM algorithm depends on an initial guess of the parameter, which may influence the uncertainty of the model. Our assumption on the level of uncertainty and the choice of the learning method significantly impact how much the generated BNs fit the learning data and how flexible the BNs respond to future inferences. We test the representational performance of the model using maximum-entropy, minimum-entropy, and EM algorithm. For EM algorithm we vary the initial parameter from 0.5 to 1. The results are presented in Table 12, ordered by the level of uncertainty. We find out that maximum-entropy (MaxE) generates the same model as EM algorithm with initial CPTs equal to 0.5 as suggested by Cowell (1999). According to Table 12, the representational performance is in decreasing order along with the level of uncertainty. It means the more uncertain a model is, the better it can represent the original stories. It is because by maximizing entropy missing values are filled with uncertain probabilities. According to the definition of compellingness, arguments with uncertain values are relatively uncompelling, thus missing arguments are correctly classified as implicit.

Table 12 Representational performance of BNs with different levels of uncertainty

Initial guess of EM or Entropy method	Hr2Hs		Dr2Ds		Hr2Ds		Hr2Ds v.s. Dr2Ds
	score	Std. deviation	score	Std. deviation	score	Std. deviation	Significance (p-value)
0.5/MaxE	0.9133	0.1094	0.8113	0.1121	0.751	0.1139	0.0114
0.5-0.75	0.8653	0.1430	0.8163	0.1446	0.7666	0.1436	0.1548
0.75-1	0.794	0.1752	0.751	0.1689	0.7103	0.1785	0.6064
MinE	0.751	0.1892	0.5176	0.1659	0.629	0.2113	2.03E-08

4.3.5.3 Weight on Individual Stories

The learning data for an individual agent is composed of all true stories in the dataset with a higher weight assigned to the individual's story. The purpose is to personalize the model so that it will most likely infer the same arguments as those in the original story while still preserving the common knowledge of the domain. How much weight we put on a speaker's story determines how biased his model is towards his subjective beliefs. An overly biased model is not flexible enough to change according to evidence, whereas an under-biased model is not able to represent the individual but simply agrees with the majority. In practice, under the constraint of favoring the speaker's own story we want to minimize the distance between the individual and the majority, thus when we vary the weight on an speaker's story, to form the learning data we select either the weight we choose or the minimum weight that satisfies the constraint, whichever is smaller.

The result of the test is shown in Table 13. The results show that 100:1 and 60:1 are reasonable ratios of weights since they are able to produce models that are diverse enough to represent their distinctive deceptive stories. It seems that the performance of 100:1 is not significantly better than that of 60:1. This is because most of the models can infer the original story as the MPE with a weight less than 60. On the other hand, models

learned from the data formed by the weight of 30 cannot incorporate personal difference appropriately.

Table 13 Representational performance of BNs with different interpretations of uncertainty

weight	Hr2Hs		Dr2Ds		Hr2Ds		Hr2Ds v.s. Dr2Ds
	score	Std. deviation	score	Std. deviation	score	Std. deviation	Significance (p-value)
Min (min, 30)	0.8853	0.1073	0.804	0.1129	0.7593	0.1177	0.0825
Min (min, 60)	0.9103	0.1081	0.8133	0.1131	0.7543	0.1165	0.0126
Min (min, 100)	0.9133	0.1094	0.8113	0.1121	0.751	0.1139	0.0114

4.3.5.4 Size of Learning Data

The learning data basically determines the context knowledge, or the "common sense" generally believed by the group of agents. The context knowledge provides a reference to the agents when inferring with unseen evidence. The more complete it is, the more possibilities an agent can think of. A larger group of agents can provide more complete context knowledge with more detail given other things being equal. Whereas an agent built based on knowledge from a small group may lose possible arguments due to the lack of detail. Consequently, the stories inferred by an agent, honest or deceptive, will not be very different from the agent's learning data. This is exactly what we see from the test results in Table 14. The results show that 30 or fewer agents are not sufficient to form a comprehensive context to build the BNs. In reality, the requirement of having more than 30 agents delivering opinions on the same topic may be infeasible. If context knowledge can be provide by domain experts we may be able to use fewer agents to achieve the same level of performance.

Table 14 Representational performance of BNs with different sizes of group

Group size	Hr2Hs		Dr2Ds		Hr2Ds		Hr2Ds v.s. Dr2Ds
	score	Std. deviation	score	Std. deviation	score	Std. deviation	Significance (p-value)
100 agents	0.9133	0.1094	0.8113	0.1121	0.751	0.1139	0.0114
30 agents	0.9016	0.1117	0.799	0.1142	0.746	0.1181	0.0603
10 agents	0.9106	0.1228	0.793	0.1216	0.7466	0.1179	0.2547

4.3.5.5 Similarity between Agents

There are both advantages and disadvantages to using similar agents. Similar agents can help generate the true stories more precisely because every agent agrees with the "common sense". However, if they are not diverse enough they cannot provide us with both sides of the story. As a result of this, the inferred stories will be overly certain and insensitive to evidence. In this test only the stories that deviate from a speaker's story by a certain amount of arguments are included in the speaker's learning data when building his BN. The result in Table 15 credits our hypothesis by showing that stories different from each other by more than 5 different arguments are necessary to build models that can represent both true and deceptive stories, but, with a larger difference with the group, an agent's performance of Hr2Hs and Dr2Ds are decreased.

Table 15 Representational performance of BNs with different similarities between agents

Different arguments #	Hr2Hs		Dr2Ds		Hr2Ds		Hr2Ds v.s. Dr2Ds
	score	Std. deviation	score	Std. deviation	score	Std. deviation	Significance (p-value)
0-5	0.9536	0.0487	0.7926	0.0923	0.7746	0.0864	0.2857
5-10	0.9073	0.1138	0.803	0.1162	0.744	0.1170	0.0318
10-	0.89	0.1649	0.7893	0.1443	0.72	0.1416	0.0164

The parameters in the learning of cognitive models that have impact on the representational performance include the accuracy of argument retrieval, the level of

uncertainty of the agents, the weight of individual's stories in learning data, the size of learning data and the similarity between agents. Below we summarize their impacts on the learning of cognitive models.

1. Automated retrieval of arguments may not be applicable to our framework since the retrieved arguments are not diverse enough to generate personalized models.
2. The level of uncertainty determines how missing data is handled during the learning of BN parameters. Maximizing the uncertainty improves the representational performance as well as personalizes the individual models.
3. The weight of an individual's story in the learning data controls the bias of the agent towards his subjective beliefs. For the abortion data, we can generate models that favor their individual stories with a weight lower than or equal to 60. We think this parameter may be specific to topic. In topics where attitudes are diverse, a higher weight of or a larger bias towards an individual's story is necessary.
4. The size of group determines the completeness of the context knowledge of the agents. The more complete the context knowledge is, the greater the number of possibilities of arguments the models can provide.
5. The similarity between agents determines whether the learned BNs cover both sides of the story, and thus influences the certainty and the flexibility of the BNs.

Although there remain problems to be addressed, learning of cognitive process from natural language can be approximated to some extent using a combination of NLP and modeling techniques as described. Improvement in the performance can be anticipated if

more advanced and relevant NLP tools are utilized. However, as we have argued, the construction of the cognitive process is not an essential part of the detection method if sufficient data is provided. To execute the detection method, it is only necessary to retrieve the semantics of arguments from the verbal content.

4.4 Detection Results and Evaluations

We performed deception detection on the two datasets (hotel reviews, and abortion data) based on the detection model in Fig. 3. For each dataset we first synthesized 100 repeats of historic data for each agent. The historic data is treated as the training data, which is then used to form correlation networks and provide information on the relations between arguments for the purpose of measuring reasoning patterns. Each story in the dataset is then treated as a repeat of testing with a distinctive set of evidence. In the testing process we assume that only the i^{th} agent at the i^{th} repeat exhibits deceptive behavior, which is represented by the deceptive stories. For the i^{th} repeat of testing, we synthesize an honest story for each agent other than the i^{th} agent by supposing that each agent had provided a story with the same evidence as the i^{th} agent. These honest stories are used to form consensus networks and calculate the inconsistency and the untruthfulness of the stories in the testing data. The reasoning patterns are evaluated based on the inconsistency and untruthfulness together with the relations between arguments believed by the respective agents. Finally, we obtain the scores of the reasoning patterns for each story in the testing data, which are classified as deception or truth through logistic regression. The entire process of the deception detection is summarized in Table 16.

Table 16 Steps of deception detection

Step1	Build $agent_i$ for the i^{th} speaker based on the true dataset.
Step2	Sample the agents to simulate several repeats of historic data.
Step3	Obtain the correlations between agents and form correlation networks. Predict the historic data and measure the prediction errors.
Step4a	For the i^{th} repeat of the true dataset, simulate honest data for the $agent_j$ where $j \neq i$ using evidence in the i^{th} true story. Form consensus network for $agent_i$ in the i^{th} repeat of the true dataset.
Step4b	For the i^{th} repeat of the deceptive dataset, simulate honest data for the $agent_j$ where $j \neq i$ using evidence in the i^{th} deceptive story. Form consensus network for $agent_i$ in the i^{th} repeat of the deceptive dataset.
Step5a	Measure inconsistency in the i^{th} true story by taking the difference between the true story and the predicted true story.
Step5b	Measure inconsistency in the i^{th} deceptive story by taking the difference between the deceptive story and the predicted deceptive story.
Step 6a	Measure untruthfulness in the i^{th} true story by comparing the true story with $agent_i$'s consensus network in the i^{th} repeat of the true dataset
Step 6b	Measure untruthfulness in the i^{th} deceptive story by comparing the deceptive story with $agent_i$'s consensus network in the i^{th} repeat of deceptive dataset
Step 7	Measure reasoning patterns of the i^{th} true story and the i^{th} deceptive story for the i^{th} agent.
Step 8	Classify the i^{th} true story and the i^{th} deceptive story as deception or truth.

4.4.1 Evaluation of Inconsistency Detection

We first evaluate the performance of inconsistency detection since we are the first to apply it to deception detection and the detection performance heavily depends on it. We will perform a pilot study using simulation data to evaluate its performance in general, followed by the evaluation on the deception datasets.

To ensure that the experimental results apply in general we simulate the knowledge base of a speaker using an existing BN and perform random test cases. The Alarm Networks (Beinlich et. al, 1989) was selected as our first test subject due to its reasonable size. Belief updating was performed to generate the reasoning results for the ease of the experiment. By perturbing the CPTs in the Alarm Network we simulated agents that are

slightly different in their conditional probabilities, which would reflect similar but not exactly the same uncertainty about knowledge. We used a perturbation value to control the noise added in the conditional probabilities. For example, if the perturbation value is 0.1, the noise to be added is within +/-0.1. In this study, 1000 repeats of inference were conducted on 100 simulated agents, each with a different set of 10 pieces of evidence, both in the training and testing processes. In the testing process, the agents generate two versions of values: true values and deceptive values. The deceptive values are simulated by simply rotating each agent's posterior probabilities. The methodology of inconsistency detection can be summarized as follows: First, we calculate the correlations between each two agents by comparing their past opinions. Next, based on the GroupLens prediction technique (Resnick et. al, 1994), we predict each agent's opinion about the current task. Finally, inconsistency will be identified if the predicted opinions are far different from the actual opinions. In practice, we allow 4 std. deviations for a normal prediction error. The effectiveness of this method has been shown in Santos and Johnson's study (2004). Here we repeat the experiment with a modified parameter setting in order to verify the results and provide a more comprehensive analysis. Table 17 shows the experimental result which echoes earlier studies. From the data we can see that the mean detection rate is around 87%, which is much higher than the human detection rate (around 55%). The false alarm rate is around 1%, which is also acceptably low.

Table 17 Statistics on the inconsistency detection rates of alarm network

Parameters	Agents=100, Repeats = 1000, Perturbation = 0.1, Evidence = 1-10, std. deviation # = 4				
True positive Rate	Max	1.0	False positive Rate	Max	0.2518
	Min	0.3770		Min	0.0
	Mean	0.8716		Mean	0.011
	Med	0.9627		Med	0.003

In addition to validating the performance on the Alarm Network, we further considered how the system performs using general BNs as testbeds. As such, we conducted the same experiment on several other BNs which are Hailfinder Network (Abramson et. al, 1996), Diabetes Network (Andreassen et. al, 1991), and Munin Network (Andreassen et. al, 1989) with increasing number of nodes and increasing complexity of structure. The structures and the detection rates of the networks can be found in Table 18. We surprisingly find out that the Diabetes network has the lowest detection rate (0.443) although its number of nodes, number of states, and number of arcs are not among the largest. We further studied the structures of the networks. One discovery was that the height of the Diabetes network is more than 100 levels while the other networks' heights are within 20 levels. Inspired by Yuan's idea (Yuan, 2007) that detection rate is largely influenced by the network's intra-dependency, which is a measure of how dependent the states' probabilities are on the evidence, we hypothesize that in the Diabetes network, nodes have the least dependence on evidence because the nodes are highly separated from one another. Experiments supported our hypothesis. Since detection rate is positively correlated to intra-dependency, which means that the detection rate increases with the increase of the intra-dependency; the low detection rate of the Diabetes Network is shown to be due to its great height. Overall, the detection method is valid on networks with moderate intra-dependencies. If the height of the network is too large, the network will be too weak to propagate the evidence to all the nodes, and thus some inconsistent information cannot be detected. For further information about intra-dependency and the experimental results please refer to Appendix A.

When detecting inconsistency from the abortion data we regard an argument as inconsistent if its prediction error deviates from the prediction errors of the training data by more than 3 std. deviations. The detection result in Table 19 shows that 20.37% of the arguments in the deceptive stories are classified as inconsistent and 7.73% of the arguments in the true stories are classified as inconsistent. Out of all 200 stories only 20 of the true stories do not exhibit any inconsistency. The inconsistency in the true stories may not be completely false alarms since deviations from the self can be due to opinion change, innovation and even mistakes. Although the inconsistency rates in the deceptive stories seem low compared with the results in the pilot study, it is reasonable because normally only part of the phrases in a deceptive story are deceptive, but unfortunately we do not have the ground truth of that fidelity.

Table 18 Structures and detection rates of hailfinder, diabetes, munin networks

Parameters	Agents=10, Repeats = 1000, Perturbation = 0.1, Times of std = 4, Evidence = 30% of total nodes				
Network	no. of Nodes	no. of States	no. of Arcs	Mean positive detection rate	Mean false detection rate
Hailfinder	56	223	66	0.8050	0.0264
Diabetes	413	4682	602	**0.4430**	**0.0144**
Munin	1041	5651	1397	0.6283	0.0216

Table 19 Performance of inconsistency detection with the abortion data

Parameters	Agents = 100, Repeats = 100, Perturbation = real, Evidence = 1-7, Times of std = 3	
	Inconsistency rate in Deceptive Story	Inconsistency rate in True Story
Max	0.8	0.766667
Min	0.033333	0
Mean	0.203667	0.077333
No. of stories with inconsistency	100	80

4.4.2 Evaluation of Deception Detection

An ideal deception detection method should be both accurate and reliable, meaning that it should successfully distinguish deception from truth and that its performance is robust to the change of environment. To evaluate the performance we compare our method with existing verbal-cue methods with respect to both accuracy and reliability. The recognized methods include bag of words, 2-gram word features, POS tags and LIWC. All the methods use supervised classification models to classify deceptive/true texts. They basically differ in the features retrieved from the texts. Bag of words models a text by a vector of the appearance of words (with or without frequency). The models based on n-gram word features calculate the probability of n-gram patterns of words showing up in an honest/deceptive story. Based on this, the probability of a text being honest and deceptive can be obtained. To extract the n-gram word frequencies, we used the SRI Language Modeling Toolkit, which provides free applications for building and applying statistical language models primarily for use in statistical tagging and segmentation. POS tags can be automatically retrieved from given texts using Stanford Log-linear Part-Of-Speech Tagger. For LIWC, which retrieves the cognitive features from a text, we use the whole of 29 variables of LIWC for the classification. In addition, it might also be interesting to see whether classification on simply the semantic arguments can be effective in deception detection with the hypothesis that deceivers tend to raise some arguments that truth tellers usually neglect. Naïve Bayes and SVM are found to be most effective classification methods in classifying deceptive/true texts based on these features.

To compare our results on the hotel reviews with the results of the conventional cueing methods, we need to synthesize the natural-language stories of the agents. What we did is

we collected all the sentences that represent each argument from the original reviews and built a pool of true arguments and a pool of deceptive arguments. To synthesize the language based on our inferred arguments, we randomly selected sentences from a truth pool or a deception pool for the corresponding arguments. Below shows a synthesized story. The order of sentences is re-arranged for the ease of comprehension.

> *Our room was really nice, and the best bed ever. We found all of the staff very helpful and prompt. But if you are looking for a hotel with a lot of restaurants around it, the hotel is far away from many attractions. Hope I will be able to stay there again in the future.*

In Table 20 and in Table 21 we show the detection performance of our method compared with existing methods. All methods were evaluated using 10-fold cross-validations. The classification rate is the percentage of stories that are correctly classified. This accounts for the detection of both the true stories and the deceptive stories. Precision, recall and F-measure are the basic measures of performance in information retrieval. In the context of classification, precision measures the fraction of positive results (data classified as deception) that are actually deceptive, and recall measures the fraction of deceptive data that are detected. Another measure that is probably more relevant in the context of classification is the area under ROC curve (AUC) (Fawcett, 2006). The ROC curve is a graphical plot of the performance of a binary classification model varied by its discrimination threshold. The graph plots the true positive rate versus the false positive rate with different thresholds. A dot under the diagonal (x=y) of the graph represents a model that generates more false positives than true positives, therefore, the bigger the

area under the curve, the better the model performs. We compare the detection methods with respect to all the measures. The highest rate of each measure is bolded and underscored. We also bolded the highest rates from only the cueing methods. For the model based on bag of words, we provided the results on the abortion data that were claimed in Mihalcea's work (2009) together with the results we produced. We can see that for both datasets our method has the highest classification rates and F-measures. We also have the highest AUC for the hotel reviews. Compared with the other methods including human detection, we have improved the performance of deception detection by 3.5% to 29.5%. The performance with the hotel reviews is significantly better than that with the abortion data due to two reasons: (i) the hotel reviews are composed of synthetic arguments which strictly follow our proposal of the reasoning processes, and (ii) the detection is easier in fact-based data versus in opinion-based data. Except for our method, for the hotel reviews, the 2-gram model has the highest classification rate and F-measure, and the bag of words model has the highest AUC. For the abortion dataset, argument classification has the highest classification rate and AUC, and the bag of words model has the highest F-measure.

Table 20 Deception detection performance with the hotel reviews

Hotel reviews		classification	precision	recall	F-measure	AUC
Classification using arguments	Naïve bayes	0.64	0.637	0.65	0.644	0.599
	SVM	0.59	0.587	0.61	0.598	0.59
Classification using bag of words	Naïve bayes	0.76	0.789	0.71	0.747	**0.875**
	SVM	0.69	0.69	0.69	0.69	0.69
Classification using 2-gram$^+$ frequency	Naïve bayes	0.77	0.787	0.74	0.763	0.87
	SVM	**0.81**	**0.798**	**0.83**	**0.814**	0.81
Classification using POS	Naïve bayes	0.565	0.592	0.42	0.491	0.6
	SVM	0. 53	0.538	0.43	0.478	0.53
Classification using LIWC	Naïve bayes	0.62	0.64	0.55	0.591	0.686
	SVM	0.665	0.651	0.71	0.679	0.665
Our method with synthesis data		**0.845**	**0.842**	**0.85**	**0.846**	**0.898**

Table 21 Deception detection performance with the abortion data

Abortion data set		classification	precision	recall	F-measure	AUC
Classification using arguments	Naïve bayes	0.65	**0.679**	0.57	0.62	**0.745**
	SVM	**0.685**	0.664	0.75	0.704	0.685
Classification using bag of words	Naïve bayes	0.70(claimed)/ 0.675	0.628	**0.86**	**0.726**	0.729
	SVM	0.675(claimed)/ 0.635	0.624	0.68	0.651	0.635
Classification using 2-gram$^+$ frequency	Naïve bayes	0.65	0.656	0.63	0.643	0.661
	SVM	0.665	0.639	0.76	0.694	0.665
Classification using POS	Naïve bayes	0.60	0.574	0.78	0.661	0.67
	SVM	0.66	0.638	0.74	0.685	0.66
Classification using LIWC	Naïve bayes	0.62	0.64	0.55	0.591	0.686
	SVM	0.68	0.676	0.69	0.683	0.68
Our method		**0.725**	**0.683**	0.84	**0.753**	0.739

In general, word-level classifications outperform most other cueing methods. It is probably because a context-sensitive approach is more accurate than a universal feature-

based approach. Our method performs even better than word-level classifications because we measure universal patterns with reference to context knowledge. However, as we have argued in Chapter 1, features retrieved by word-level classifications are heavily domain dependent. In (Mihalcea et al., 2009), the authors attempted to classify deceptive stories of one topic using the discriminative features obtained from another topic. The performance was found to be no better than the human detection rate. We also evaluated this finding using the databases we have. We first identified the most discriminative words by calculating the dominance score of each word using a modified version of the formula proposed by Mihalcea et. al. (2009):

$$Dominance_D(W) = \frac{Coverage_D(W)}{Coverage_T(W)}$$

where $Coverage_T(W)$ denotes the word coverage in the true corpus *T*, which is defined as the percentage of words from *T* being word *W*, and the dominance score $Dominance_D(W)$ of word *W* is the ratio between its coverage in the deceptive corpus with respect to its coverage in the true corpus. The word coverage is obtained by:

$$Coverage_D(W) = \frac{Frequency_D(W)}{Size_D}$$

where $Frequency_D(W)$ represents the total number of occurrences of word *W* inside the corpus *D*, and $Size_D$ represents the total size of corpus *D*. A dominance score close to 1 indicates a similar distribution in the true corpus and the deceptive corpus. A score significantly larger than 1 indicates that the word is dominant in the deceptive corpus, and thus is likely to be a discriminative word for deception. On the other hand, a score significantly smaller than 1 indicates dominance in the true corpus.

We collected the 50 most discriminative words that are dominant in both the true stories and the deceptive stories. We found that around half of them are domain-specific (such as "room" in the hotel reviews). If we compare the discriminative words from the hotel reviews with those from the abortion data, only 4% of the discriminative words in the deceptive stories are the same and 0% of the discriminative words in the true stories are the same. An even more shocking finding is that some discriminative words for deceivers in one dataset can be dominant in the true stories in another dataset. For example, "definite" is a discriminative indicator of deception in the hotel reviews but becomes an indicator of truth in the abortion data. In total, 6% of the words that are dominant in the true abortion stories are actually dominant in the deceptive hotel reviews, and 2% the words that are dominant in the deceptive abortion stories are dominant in the true hotel reviews. The reason might be because people show different cognitive features when arguing about different topics. For example, people defending their own values may sound more emotional and those describing experiences may sound neutral. The gap is not only due to topical difference but also caused by individual difference. This can be observed by the discriminative words within the same database. We partition the true stories of the hotel reviews into 2 distinct parts, performed deception detection on the sub-datasets, and compared their discriminative words. Only 22% of the dominant words in the true stories are the same, therefore, we concluded that word features of deception are different across people and across topics. Classifiers trained on one dataset cannot be easily applied to new topics and new speakers. Another problem with respect to the reliability of word-cue methods is the possibility to escape detection by avoiding the discriminative words. To test this hypothesis we performed classifications based on the

bag of words model again after replacing the discriminative words with their synonyms. We respectively changed 1%, 2%, 3%, 4% and 5% of the most discriminative words in each test and plotted the detection performance in Fig. 10. For both classifiers, the detection performance decreases as more discriminative words are replaced by synonyms. Compared with verbal cueing, our method is more reliable because i. it can work independently of the topic domain as we capture universal reasoning patterns that all deceivers exhibit given texts with sufficient levels of reasoning, and ii. by capturing the reasoning process, it is robust to the careful craft of wording and the change of communication factors because of the nature of deceit and the deceivers' lack of true knowledge.

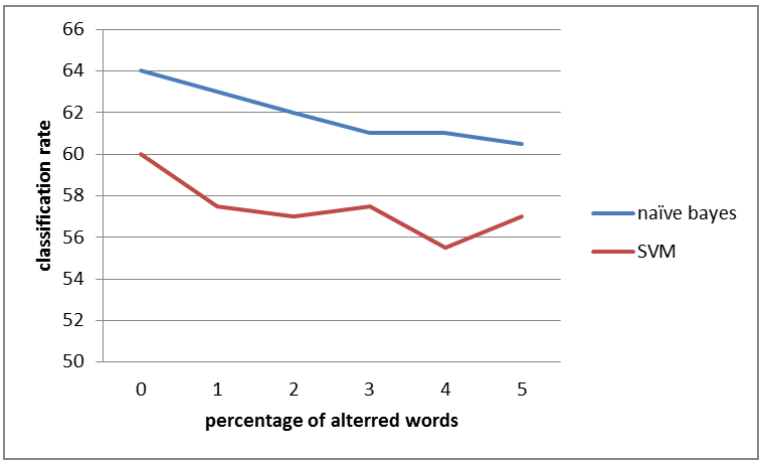


Figure 10 Classification rates with the abortion data using bag of words by replacing discriminative words with synonyms

We also observe from Table 21 that argument classification outperforms most other methods only in the abortion data. We notice that the true stories are dominated by opinions supporting abortion and the deceptive stories are dominated by opinions against abortion. Thus it occurs to us that the performance might be related to the population of

different attitudes on the topic. We looked at the discriminative arguments that have a dominance score smaller than 0.5 (dominating in the true stories) or bigger than 2 (dominating in the deceptive stories) and evaluated their relationships with the conclusions. The discriminative arguments are presented in Table 22. In the table, a postfix of 0 denotes the negative state of an argument and a postfix of 1 denotes the positive state. It turns out that among all 6 arguments that are dominating in the true stories, 5 of them (unbolded) support abortion; and among the 12 arguments that are dominating in the deceptive stories, 11 of them (unbolded) are against abortion, therefore, the discriminative arguments are heavily biased by the imbalanced size of stories with different attitudes. With the positive attitude dominating the true stories and the negative attitude dominating the deceptive stories the bias in the population facilitates the detection because rather than classifying deception the method is classifying attitude. However, if different attitudes in the datasets are balanced, the classification using arguments may not be as successful. The imbalance in attitudes not only interferes with the classification on arguments, but also may bring about potential problems in word-level classifications, that is to say, the words with discriminative power as proposed by existing research could possibly be words that support certain attitudes instead of words that deceivers tend to use.

Table 22 Discriminative words for the deceptive stories and the true stories in the abortion data

Classification	Discriminative words
True	**time1,** child0, circumst1, legal1, carri0, god0
Deceptive	govern1, carri1, circumst0, legal0, life0, right0, world1, rape0, **popul1**, birth0, god1, care0

Our intent-based method achieves the best detection performance so far. We explored whether a combination approach using both our method and the bag of words model can

further improve the performance. To this end we combined the feature vector of the reasoning patterns with the word features from the bag of words model, and built classifiers using Naïve bayes, SVM and logistic regression. The results in Table 23 and Table 24 show that in both datasets the Naïve Bayes classifer improved the performance with respect to all measures. Compared with all other methods, it achieved an improvement of 5% to 31%, but the improvement against our intent-based model is at the cost of the complexity of a Naïve bayes classifier.

Table 23 Deception detection performance with the hotel reviews using combination approach

Hotel reviews		classification	precision	recall	F-measure	AUC
Classification on reasoning feature + bag of words	Naïve bayes	**0.86**	**0.919**	0.79	**0.849**	**0.946**
	SVM	0.85	**0.85**	0.85	**0.85**	0.85
	logicstic	0.79	0.764	0.84	0.8	0.856

Table 24 Deception detection performance with the abortion data using combination approach

Abortion data set		classification	precision	recall	F-measure	AUC
Classification on reasoning feature + bag of words	Naïve bayes	**0.77**	**0.729**	0.86	**0.789**	**0.829**
	SVM	0.645	0.628	0.71	0.667	0.645
	logicstic	0.535	0.535	0.54	0.537	0.549

4.5 Explanation of Detection

Detected deception can be explained by examining how the deception was formed and how the deception was detected. The former way explains the tactics of the deceiver while the later explains the strategy of the detector. Our method is able to provide information on both. We will illustrate the analysis using an example in the synthesized hotel reviews. The example we present contains a deceptive story and an honest story of a deceiver. Below (Table 25) lists the arguments in the stories. The arguments that deviate from the honest story are in bold.

Table 25 A synthesized story of hotel reviews

Argument	Honest	Deceptive
tv	-1	-1
fitness	-1	-1
charge_for_extra	0	0
service	**1**	**0**
checkin_checkout	-1	-1
front_desk	**1**	**-1**
housekeeping	-1	-1
staff_respond	**-1**	**0**
location	1	1
access_to_shopping_and_eating	-1	-1
access_to_attractions	**1**	**-1**
michigan_ave	0	0
stay	**0**	**-1**
view	-1	-1
dining	-1	-1
design	-1	-1
facilities_condition	-1	-1
booking_agent	-1	-1
superior	**-1**	**0**
positive_reviews	-1	-1
return	**1**	**0**
recommend	**1**	**0**
functions	0	0
room	**1**	**0**
conclusion	**1**	**0**
bed	**1**	**-1**
bathroom	-1	-1
room_type	-1	-1
rate	-1	-1
parking	-1	-1
wifi	0	0

How did the deceiver form the deceptive story? We can make a reasonable guess by comparing his deceptive arguments with their honest versions. The speaker's purpose was to provide negative comments on the hotel while actually his experience in the hotel was positive. To this end, by distorting his opinions on the room and the service he claimed that this hotel was not as good as others and that he would neither come back nor

recommend it to anyone else. To prove that the hotel is worse than other hotels, he concealed the fact that the hotel is within easy access to attractions and fabricated a negative experience in which the hotel staff did not respond to his request appropriately. For the convincingness of his arguments he also concealed the positive comments on the friendly front desk service and the good quality of the bed. Note that this explanation may not be the ground truth since we cannot guarantee that all deviated arguments are deceptive.

By utilizing our detection model, a similar explanation which provides more information than the above guess about the deceiver's manipulations can be automatically obtained based on the following analysis. First of all, we can restore the story by predicting the honest arguments using GroupLens method. Through comparing the deceptive arguments with their honest probabilities we can find out how the deceiver manipulates the arguments as in Table 26. A manipulation from an uncertain state to a certain state is a fabrication of facts; reversing the state of an argument is a distortion of facts; and increasing the uncertainty of an argument is a concealment of facts. We notice that our method identifies not only the arguments that were possibly fabricated by the author but also those that were possibly hidden by the author (such as "stay"). Statements with hidden information are defined as half-truth (Crystal, 2003), which is also an intentional deception. A deceiver may indicate false information from a half-true statement such as "he is starving the cat" by omitting the fact that the cat is receiving a treatment. The ability to restore hidden information is a valuable but uncommon capability. Then we build a relational graph (Fig. 11) of all manipulated arguments based on their mutual

dependence calculated using the method in Chapter 3.2.2.2. Arguments with small dependences on all other arguments are omitted since their inconsistencies are not driven by the goal (conclusion) of the deceiver. This graph intuitively illustrates the flow of the manipulation, based on which we can derive a similar explanation with our earlier guess. This analysis shows how the false information was conveyed to the listener in a convincing manner by pinpointing the deceptive arguments and depicting the flow of possible deception.

Table 26 Manipulated arguments and predicted truth for the purpose of deception explanation

Argument	Predicted honest	Actual deceptive	Manipulation
service	0.926693	0	Distort
staff_respond	0.432515	0	Fabricate
stay	0.10014	0.5	Conceal
superior	0.451421	0	Fabricate
return	0.890742	0	Distort
room	0.867221	0	Distort
conclusion	0.87992	0	Distort

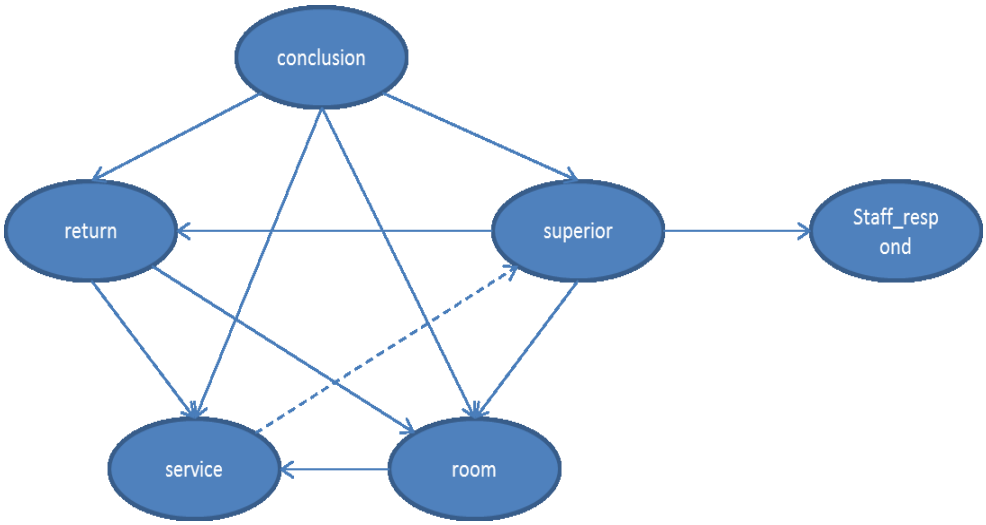


Figure 11 Relational graph of manipulated arguments

As a detector our model judges based on the reasoning patterns that were discussed in Chapter 3.2.2. Each pattern is a feature of deceptive reasoning, thus our model can

provide information about which aspect of the story indicates deception and how strongly each indicator supports the detection. Using the same example we analyze why the story is deceptive by looking at the three patterns. Table 27 lists the mutual dependence between the inconsistent arguments. Assuming that a dependence value smaller than 0 indicates independence, we found that "stay" and "bed" are independent of other arguments. This means that their inconsistencies are not caused by the impact of the conclusion, which reduced their suspicion of being deceptive. According to our calculation of functionality (Table 28), all 7 inconsistent arguments are more functional to the conclusion compared with their honest versions. Furthermore, the comparison between their inconsistencies and untruthfulness in Table 29 shows that all of them are truthful (below 3 std. deviations). Overall, 5 of the 7 inconsistent arguments exhibited all three patterns and 2 of them activated two patterns (Table 30). By thresholding the pattern scores and the number of activated patterns human detectors can classify deception based on their own knowledge and experience.

Table 27 Mutual dependence between inconsistent arguments

Dependence	service	staff_respond	stay	superior	return	room	conclusion	bed
service	n/a	-0.015	-0.535	0.077	0.187	0.161	0.167	-0.296
staff_respond	-0.015	n/a	-1.140	0.209	-0.005	-0.030	-0.035	-0.719
stay	-0.535	-1.140	n/a	-0.673	-0.433	-0.434	-0.426	-0.587
superior	0.077	0.209	-0.673	n/a	0.172	0.171	0.163	-0.373
return	0.187	-0.005	-0.433	0.172	n/a	0.143	0.163	-0.330
room	0.161	-0.030	-0.434	0.171	0.143	n/a	0.163	-0.349
conclusion	0.167	-0.035	-0.426	0.163	0.163	0.163	n/a	-0.334
bed	-0.296	-0.719	-0.587	-0.373	-0.330	-0.349	-0.334	n/a

Table 28 Functionality of inconsistent arguments

	service	staff_respond	stay	superior	return	room	bed
functionality	1	1	1	1	1	1	1

Table 29 Comparison of the inconsistency and the untruthfulness

	service	staff_respond	stay	superior	return	room	bed
Inconsistency	6.4576	4.2657	4.3688	4.3601	6.5313	5.5116	5.3940
untruthfulness	0.9761	1.0525	0.0953	0.4409	0.4844	0.7968	0.0622

Table 30 Activation of deceptive patterns

	service	staff_respond	stay	superior	return	room	bed
Activated pattern #	3	3	2	3	3	3	2

4.6 Misinformation

As we have mentioned in Chapter 1, malicious intent is a key component of deception. It determines whether an unexpected deviation is an unintentional error or an intentional act of misleading. It is necessary to distinguish deception from unintentional errors because deceivers bring long-term damage with oriented targets while error maker is not aware of the falsity/deviation in his communication and does not present the false/deviated information regularly. In this dissertation, unintentional errors are referred to as misinformation. It is a challenging task to discriminate deception and misinformation using either computational or psychological methods because their effects are very similar. Our model is able to distinguish deception from misinformation because the reasoning patterns are derived from the unique cognitive process of deceivers. We have verified that deception and random manipulations are distinguishable by the patterns using simulation data in Chapter 3.2.2.1. The purpose of this section was to test our detection method using real world misinformation data, but unfortunately none of the existing deception datasets contains misinformation data. To evaluate the performance of our framework we synthesized misinformation data for both datasets. We consider three major types of misinformation: misunderstanding, wrong assumptions, and opinion change. Misunderstanding is defined as “a form of understanding that is partially or

totally deviant from what the speaker intended to communicate" (Weigand, 1999). Milroy (1984) describes misunderstanding as "the disparity between the speaker's and the hearer's semantic analysis of a given utterance". That is to say, the speaker's attitude on an argument is misinterpreted to another attitude. Misunderstanding occurs after the speaker has presented or formed his arguments. Moreover, the interpretation of one argument is not shown to be related to the interpretation of other arguments. Thus, to simulate misinformation we perturbed the probabilities of each argument by adding random noise. Assumptions are represented by evidence in our cognitive model. In each repeat of reasoning a set of evidence is fed to the BN to constrain the reasoning result. Wrong assumptions mean that the speaker selected wrong evidence before reaching conclusions so wrong assumptions can be simulated as randomly setting evidence before conducting reasoning. However, for the realness of the synthesis data the size of random evidence is constrained by the common sizes of evidence in the original stories. Opinion change indicates that one's actual opinion is different from what was believed in the old knowledge. It is obvious that changed opinion is derived from valid and honest reasoning, thus the cause of a changed opinion is fundamentally a change in the knowledge base. In our cognitive models knowledge bases are represented by BNs, therefore, opinion change is simulated by perturbing the CPTs of an agent's BN.

Although it contains errors, misinformation is derived from valid and honest reasoning, thus, truth and the three types of misinformation are expected to differ from deception in terms of the reasoning patterns. The difference may not exist in the occurrence of the patterns but in the strength of each pattern. To provide an intuitive demonstration of the

difference we first present the scores of patterns for each type of reasoning. Our expectations of the strengths of patterns are listed in Table 31, in which ✓ denotes a strong pattern, Δ denotes a weak pattern, and × denotes no pattern. Propagation of manipulation happens only when inconsistencies are formed by natural reasoning which includes wrong assumptions, opinion change, truth and deception. However, inconsistencies in truth are mostly false alarms. We expect the amount of inconsistencies to be small and the pattern of propagation to be weak. Although opinion change follows natural reasoning, our estimate of argument relations is based on a speaker's old knowledge, and thus the pattern should be weak. Misunderstanding does not show this pattern because its error is caused by random noise. Functionality means that the arguments are manipulated to appear more supportive to the conclusion. Deception obviously shows the pattern. Wrong assumptions may or may not show the pattern depending on how strongly the wrong assumptions support the conclusion. Opinion change and truth may or may not show the pattern depending on the parameters of the BNs. Misunderstood arguments are usually not more functional than the original arguments because the conclusion was derived from the original arguments. Lastly, in deception inconsistent arguments are usually convincing while consistent arguments are not. This happens neither in truth nor in opinion change because unconvincing arguments are very rare in truthful reasoning. It does not happen in misunderstanding because arguments are manipulated randomly, and manipulated arguments are unlikely to agree with the arguments from truth tellers. In wrong assumptions we need to consider this pattern under two conditions. When the wrong assumptions are not strong enough to influence the conclusion the pattern will not appear because arguments manipulated by

the wrong assumptions are different from truth tellers' corresponding arguments. When the wrong assumption is strong enough to influence the conclusion arguments manipulated by the wrong assumptions will be more convincing than unrelated arguments. However, the wrong assumptions themselves cannot be convincing because it is not truthful. The first condition is more likely to happen in opinion-based datasets while the second condition appears more in fact-based datasets, therefore, we can expect the wrong assumptions in the abortion data to not show the pattern but those in the hotel reviews to show a weak pattern. According to our expectations it seems that among all misinformation, misunderstanding is the most likely to be identified as non-deception and wrong assumption is the most likely to be identified as deception. This is reasonable because deception, according to our proposal of the reasoning process, is a special type of wrong assumption in which the assumption is the conclusion.

Table 31 Expectations of the strengths of patterns for different types of reasoning

	Propagation of manipulation	Functionality	Association of untruthfulness & inconsistency
Misunderstanding	×	×	×
Wrong assumptions	✓	Δ	×(opinion)/ Δ (fact)
Opinion Change	Δ	Δ	×
Deception	✓	✓	✓
Truth	Δ	Δ	×

The testbed of this experiment is composed of 100 true stories, 100 misunderstood stories, 100 opinion changed stories, 100 stories with wrong assumptions and 100 deceptive stories. The purpose is to discriminate deception from all non-deceptions. To ensure that the evaluations for different types of reasoning are comparable we adjust the perturbation noise in each type of misinformation such that they produce similar numbers of inconsistent arguments. The scores of patterns for the hotel reviews are listed in Table 32,

and those for the abortion data are listed in Table 33. For the hotel reviews the scores of propagation of manipulation meet with our expectations, as do the scores of association of untruthfulness and inconsistency. The scores of functionality generally agree with expectations except that the score for true stories is slightly lower than expected. Nevertheless, it is not unreasonable because truth tellers tend to make conflicting arguments according to cognitive studies (DePaulo et al., 2003). For the abortion data the scores of functionality and the scores of association of untruthfulness and inconsistency agree with the expectations. The score of propagation of manipulation for deceptive stories is lower than expected. Our guess is that this measure strongly depends on the reasoning process of deceivers. The proposed cognitive model is not claimed to sufficiently cover every aspect of deceptive reasoning, and thus the manipulations on the arguments may not be completely caused by the impact from conclusion. That is to say, manipulations caused by other cognitive processes could interrupt the measure of contributions from the conclusion, and consequently introduce error to the score of the pattern.

After obtaining the scores of patterns for each story we use a 10-fold logistic regression to classify the stories to deception (positive), and non-deception (negative). Data is classified correctly if deception is classified as positive and truth, misunderstanding, wrong assumptions or opinion change is classified as negative. Due to the imbalance of the data labels we weight deceptive data 4 times as non-deceptive data. Since misinformation is synthesized to the level of semantic arguments we only compare our performance with that of argument classification. In Table 34 and Table 35 we present

our detection results for the hotel reviews and the abortion data, together with the results classified by arguments. As we can see, our method performs significantly better than argument classification under all measures. The performance for the hotel reviews is slightly better than the performance for the abortion data.

Table 32 Scores of patterns for different types of reasoning with the hotel reviews

Hotel reviews	Propagation of manipulation	Functionality	Association of untruthfulness & inconsistency
Misunderstanding	-0.27152	0.226776	0.144867
Wrong assumptions	0.390164	0.620128	0.392234
Opinion Change	-0.02896	0.270487	0.289683
Deception	0.647718	0.812537	0.797543
Truth	0.52567	0.190641	0.313

Table 33 Scores of patterns for different types of reasoning with the abortion data

Abortion data	Propagation of manipulation	Functionality	Association of untruthfulness & inconsistency
Misunderstanding	-0.20424	0.123909	0.217961
Wrong assumptions	0.231692	0.3075	0.195
Opinion Change	-0.0064	0.22135	0.451431
Deception	-0.04997	0.453633	0.811495
Truth	0.05208	0.251424	0.48715

Table 34 Deception/non-deception classification performance with the hotel reviews

Hotel data set		classification	precision	recall	F-measure	AUC
Classification of arguments	Naïve bayes	0.65875	0.668	0.63	0.649	0.706
	SVM	0.695	0.682	0.73	0.705	0.695
Our method on synthesis data		**0.84875**	**0.841**	**0.86**	**0.85**	**0.902**

Table 35 Deception/non-deception classification performance with the abortion data

Abortion data set		classification	precision	recall	F-measure	AUC
Classification of arguments	Naïve bayes	0.70625	0.738	0.64	0.685	0.773
	SVM	0.6675	0.752	0.5	0.601	0.668
Our method		**0.81875**	**0.784**	**0.88**	**0.829**	**0.858**

We suspect that in the above experiment misinformation stories might have similar arguments as true stories given small perturbation noise because misinformation stories were synthesized based on the evidence in the true stories, a factor which may facilitate the classification. To address this problem we synthesized misinformation again based on the evidence in the deceptive stories. In this case the misinformation stories could be very different from the true stories and the classification rate can be expected to decrease. The experimental results are listed in Table 36 and Table 37. We observe that the performance for the abortion data is degraded slightly. Since the abortion data is insensitive to evidence we estimate that the decrease in performance is due to the perturbation in misinformation. On the other hand, the performance for the hotel reviews does not decrease at all. Thus we believe that our method is able to detect deception from unseen data by accurately retrieving the universal features of deceptive reasoning.

Table 36 Deception/non-deception classification performance using deceptive story evidence with the hotel reviews

Hotel reviews		classification	precision	recall	F-measure	AUC
Classification of arguments	Naïve bayes	0.6525	0.66	0.63	0.645	0.697
	SVM	0.705	0.667	0.82	0.735	0.705
Our method on synthesis data		**0.86375**	**0.846**	**0.89**	**0.867**	**0.908**

Table 37 Deception/non-deception classification performance using deceptive story evidence with the abortion data

Abortion		classification	precision	recall	F-measure	AUC
Classification of arguments	Naïve bayes	0.69625	0.721	0.64	0.678	0.771
	SVM	0.67875	0.754	0.53	0.623	0.679
Our method		**0.78625**	**0.754**	**0.85**	**0.799**	**0.833**

4.7 Understanding the Data

Both datasets contain 100 pairs of true stories and deceptive stories, each from a rational speaker. We wonder why the detection performs differently in the two datasets besides the level of facts involved in the stories, or more generally, what factors of a dataset facilitate the detection and what factors impede the detection. Our purpose in this section is to evaluate the behavior of the deception detection model more thoroughly by investigating the datasets. We propose to investigate the datasets from the following aspects: the features with regards to the group of agents, the features with regards to the individual agents and the features with regards to the stories. The experiment was partly performed on the abortion data because of the availability of paired stories, and partly performed on simulation data in order to evaluate the direct influence from the reasoning processes.

4.7.1 With regards to the Group

The group of agents basically determines the correlation networks, thus, the major impact of factors of the group focuses on the detection of inconsistency. With regards to the group of agents we consider the group size, the similarity of agents and the number of deceivers.

4.7.1.1 Group Size

With a larger group, we may expect the decision making process of inconsistency detection to be influenced by more participants. To evaluate this influence we performed inconsistency detection by varying the number of agents from 3 to 100. In order to have a general understanding of this parameter, we tested on the simulation agents: the perturbed

Alarm Networks. Table 38 displays the means of true positive and false positive rates. As the agent number increases the detection performance improves, but the improvement is not very explicit, and no evidence shows that the agent number influences the false positive rate. The results indicate that when more truth tellers are involved in the detection we can anticipate the behavior of the agents slightly better.

Table 38 Detection performance with the size of group

Inconsistency detection	Repeats = 100, Evidence = 10, std. deviations # = 4, Perturbation = 0.2			
Agents	3	10	30	100
True positive	0.838475	0.848329	0.86418	0.856521
False positive	0.0163721	0.0142754	0.0154445	0.0157893

Since the influence of the group size on inconsistency detection is minor, and the detection of deception does not strongly depend on the rate of inconsistency, we can expect its influence on the detection of deception to be negligible.

4.7.1.2 Similarity between Agents

Since the detection of inconsistency is based on the assumption that agents are highly correlated, by varying the difference between the agents we can observe how sensitive the system is to this assumption. If the agents have similar knowledge they tend to form similar opinions and agree with each other. Consequently the inconsistency appears to be the more obvious. The difference in agents' stories is caused by their distinctive knowledge bases, therefore, to evaluate the sensitivity of inconsistency detection to agents' similarity, we vary the perturbation value added to the CPTs of each agent simulated by the Alarm Network. Table 39 displays the means of true positive and false positive rates. The perturbation value is inversely proportional to the true positive rate

because the more correlated the agents are, the more obvious the inconsistency appears to be. Since a high correlation leads to a high detection rate it will also cause a high false positive rate.

Table 39 Detection performance with the similarity between agents

Inconsistency detection	Repeats = 100, Evidence = 10, std. deviations # = 4, Perturbation = 0.2			
Pert. value	0.1	0.2	0.3	0.4
True positive	0.932075	0.856521	0.810289	0.759478
False positive	0.014392	0.015789	0.021366	0.022349

This result caught our attention because we do not want the detection to be more difficult on deceivers who are not correlated with others. In other words, we do not want deceivers to escape detection because they are unpredictable, therefore, we evaluated the deception detection performance with respect to agents with different coherences with the group. Specifically, we performed ANOVA to test the difference in the coherences between two sets of deceivers, one set contained only detected deceivers and the other set contained undetected deceivers. The coherence was calculated in two ways: in one, we calculated the average similarity between an agent and the group; in the other, we calculated the maximum similarity between an agent and the group. The experiment shows that the average similarities are not significantly different between detected and undetected agents, but the maximum similarities of detected agents are significantly larger than the maximum similarities of undetected agents. These results mean that being an outlier or being unexpectable to the group does not make a deceiver more difficult to be detected as long as there is at least one acquaintance who can anticipate the deceiver's opinions very well.

4.7.1.3 Number of Deceivers

Up to now, all the experiments we conducted contained only one deceiver in each repeat no matter how many agents are in the group. However, in reality, we may face the situation that more than one deceiver is working or even cooperating wither other deceivers to mislead the listener. Taking this into consideration we studied the performance of the model in detecting inconsistencies of several deceivers. Likewise we use Alarm Networks to represent the agents in this experiment. We adjusted the proportion of agents being deceivers while changing the total number of agents at the same time. The true positive rates are shown in Table 40. As we can see from Table 40, when half or more of the agents are honest the detection rates are above 67%, which is still relatively high compared with human detection ability. However, as soon as the majority of the agents act as deceivers, our detection rates drop rapidly. This result agrees with our real world experience that if the majority is lying, it is hard for the listener to tell the truth. Figure 12 shows the plotted detection rate against the proportion of agents as deceivers. The three lines represent the systems with different number of agents. We observe from the figure that the detection rate is inversely proportional to both the proportion of agents as deceivers and the total number of agents. However, the impact from the number of agents is relatively small. Therefore, it is more critical to make sure that the proportion of benevolent agents is high rather than to have a large number of benevolent agents for the purpose of detecting deception effectively.

Besides the influence on inconsistency detection, the number of deceivers also has a significant impact on the measure of untruthfulness because by dominating the group the

deceivers can distort the truth. We can expect that with more deceivers in the group, it is more difficult to detect deception.

Table 40 Inconsistency detection rate of adjusting the number of agents together with that of deceivers

No. of agents \ proportion of agents being deceiver	10%	30%	50%	70%	90%
3	NA	0.8538	NA	0.6642	NA
10	0.8728	0.8362	0.6783	**0.4680**	**0.1935**
30	0.8654	0.8045	0.6979	**0.4880**	**0.1815**
100	0.8502	0.7864	0.6668	**0.4515**	**0.1396**

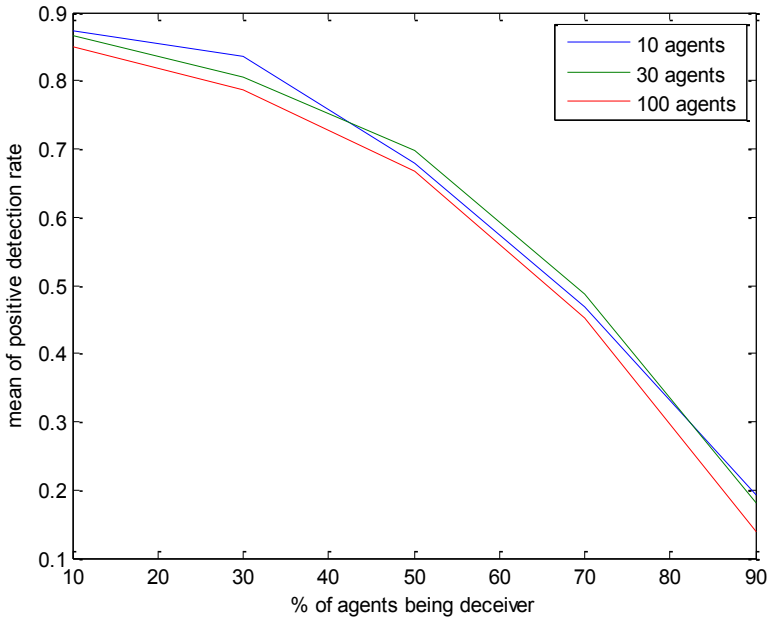


Figure 12 Plot of detection rate against the proportion of agents being deceivers

4.7.2 With regards to the Agent

Next, we study the features with regards to the individual agents. The parameters include the change in agents' arguments and the level of reasoning involved in the stories.

4.7.2.1 Change in Agents' Arguments

Since opinion-based arguments are reluctant to change with evidence, we can expect one's opinions to be similar under different evidence, and any deviation in his arguments can indicate inconsistency, thus, the difference between an agent's true story and

deceptive story can provide an estimation of the ground truth of inconsistent arguments. To verify that our model is able to catch inconsistency we measure the similarity between one's true story and deceptive story and find its correlation with the inconsistency detection rate of the deceptive story. We found that the similarity between an agent's deceptive story and true story has a large negative correlation (-0.41233) with the detection rate. It credits our hypothesis that inconsistencies can be accurately identified through our method. To evaluate the impact of inconsistency on deception detection we calculated the correlation between the similarity and the deception score obtained from the classifier of logistic regression. It turns out that the correlation (-0.25952), although not as high as with the inconsistency detection, is significantly large at 0.005 level according to the statistics on correlation significance with 100 samples (Table 41). These results indicate that the less inconsistency a deceiver demonstrates, the more difficult it is to detect deception.

Table 41 Statistics on correlation significance with 5 and 100 samples

One-tail probabilities	0.05	0.025	0.005	0.0005
Correlation for 100 samples (absolute value)	0.165	0.197	0.256	0.324
Correlation for 5 samples(absolute value)	0.805	0.878	0.959	0.99

It cannot be asserted that insisting on the same opinion can escape detection because the detection is not solely based on the distance between an agent's current opinion and historical opinion. We see this by observing the distance between an agent's training data (historical benevolent stories) and each type of testing data (stories with different types of reasoning) together with the detection performance for each type of testing data. During the synthesis of misinformation data in Chapter 4.6, we already ensured that all types of misinformation data generate similar numbers of inconsistent arguments as the deceptive

data. Thus we only look at the rates of deception detection, which are listed in Table 42. Obviously, the difference between the training data and the test data does not correlate with the detection rates (the correlation value 0.415041 is not significant with 5 samples according to Table 41). This means that the detection performance is not sensitive to the change in one's story but is influenced by the type of reasoning and the inconsistency of the speaker.

Table 42 Comparison of change in agents' arguments with detection performance

	Truth	Deception	Change	Wrong	Misunderstand
Difference in arguments with history data	1256	2377	2823	1005	1782
Detection rate	0.41	0.87	0.33	0.16	0.04

4.7.2.2 Level of Reasoning

Deception detection is more meaningful if a higher level of reasoning is involved in the stories. Deception formed by complex reasoning is more challenging, and blatant lies that simply claim a falsified conclusion without considering the consistency of the entire story or stories in which conclusions were not reached by logically relate arguments are not common in serious deceptions, therefore, in our experiments we ensure that there are sufficient levels of reasoning in the stories. It occurs to us that the hotel reviews may involve less reasoning compared with the abortion data because opinion-based stories are usually derived by competing arguments but fact-based stories are derived by memories or experiences. To verify the hypothesis, we need to measure the level of reasoning involved in each story. According to Wilhelm (2005), a major determinant of the difficulty of a reasoning task is the number of mental models that are compatible with the assumptions. The assumption "A is left of B. B is left of C. C is left of D. D is left of E." can be easily integrated into one mental model, in which the entities are arrayed as "A B

C D E" from left to right, but the assumption "A is left of B. B is left of C. C is left of E. D is left of E." calls for the construction of two possible mental models ("A B C D E" and "A B C E D"). In our cognitive model based on Bayesian inference, "mental models" refer to the possible options of arguments derived from a set of evidence through belief revision. The more options of arguments an agent can select to believe in, the more complicated the inference is. Based on this idea, we propose that reasoning can be measured by the number of hypotheses derived from a knowledge base with a set of evidence. Given this definition, a reasoning task is more complicated if

1. There are more arcs in a BN (meaning arguments are more logically related)
2. There are more nodes in a BN (meaning more details and aspects are covered)
3. There is less evidence (More evidence can constrain the number of possible hypotheses, and thus makes the reasoning task easier.)

We measured the level of reasoning in the abortion dataset, the hotel reviews and the "my best friend" dataset which describes the speaker's best friend with arguments partially based on facts and partially based on opinions. Since we manually constrain the number of nodes and select evidence, the level of reasoning can be estimated by the number of arcs in the BNs provided that the BNs have similar size. In addition to the number of arcs, we performed analysis based on basic graph theories and compared the results of different datasets. The measures that are significantly different between fact-based reasoning and opinion-based reasoning are presented in Table 43. As we have expected, there are more arcs in the BNs of the abortion data than those of the hotel reviews and the best friend data. We also observe that although there are more arguments in the fact-

based datasets, the arguments are less connected, thus, there is a lower level of reasoning in the hotel reviews and the best friend data even without considering the evidence. To ensure that the level of reasoning in the hotel reviews is comparable with that in the abortion data, we allow a higher threshold of correlation to connect nodes during the construction of BN structure. The measures of the new BN (denoted as Hotel review*) are now comparable to those of the abortion data. The hotel reviews used in the deception detection as described in Chapter 4 were synthesized based on the new BN with higher level of reasoning. Most of the measures in Table 43 are related to the number of arcs but we notice that with similar number of arcs the Hotel reviews* have fewer and bigger maximum cliques than the abortion data. Although not directly related with the level of reasoning, we prefer big numbers of small cliques instead of small numbers of large cliques because BNs with small cliques are more hierarchically organized whereas in BNs with big cliques each argument is related with each other argument. For the same reason we also prefer smaller centrality closeness given similar number of arcs and similar sizes of cliques since in a hierarchical graph nodes are relatively farther apart.

Table 43 Basic graph theory analysis on the BNs of different datasets

Dataset	Abortion data	Best friend	Hotel review	Hotel review*
Arguments # (average)	5.29	6.302703	7.96	10.91042
Arcs #	48	39	39	46
Degree (average)	3.2	2.5161	2.5161	2.9677
Max cut	42	36	35	43
Size of clique (average)	3.2000	2.5484	2.5484	3.0323
Clique number	2	2	2	3
Maximal cliques #	48	40	40	46
Cliques containing each vertex # (average)	4.133333	2.451613	2.419355	3.032258
Centrality closeness	0.3645	0.278581	0.295689	0.361191
Density	16/145=0.1103	13/155=0.0839	13/155=0.0839	10/93=0.1075
Dominating set size	7	11	10	8

4.7.3 With regards to the Story

Parameters with regards to the story include the size of historic data, the level of noise in historic data, the level of noise in testing data, and the size of evidence.

4.7.3.1 Size of Historic Data

Historic data is mostly used to calculate the correlation between agents which are basically used in the detection of inconsistency, thus, in this test we detect inconsistency from the simulation agent "Alarm Network" by varying the number of repeats in the training data. Table 44 shows the experimental results. Surprisingly, the results do not demonstrate a significant influence of the size of historic data on the true positive or false positive rate. Likewise, we do not expect a significant influence on deception detection either.

Table 44 Inconsistency detection performance with the number of repeats

Inconsistency detection	Agents = 10, Evidence = 10, std. deviations # = 4, Perturbation = 0.2			
Repeats	*10*	*100*	*1000*	*10000*
True positive	0.887225	0.848329	0.854309	0.860704
False positive	0.109151	0.0142754	0.0107823	0.0111045

4.7.3.2 Noise in Historic Data

During a communication, speakers sometimes are not able to precisely deliver their inferred arguments, but the deviation should not be big enough to change their opinions, therefore to simulate realistic historic stories, random noise was added to each agent's threshold of the compellingness of arguments. This means that speakers may mistakenly present arguments that they do not believe to be compelling and ignore arguments that they believe to be strong. We expect that the level of noise in the historic data may influence the calculation of agent correlations and result in the degradation of inconsistency detection. To this end we perform inconsistency detection on the abortion data by simulating historic data with different levels of noise. We are not able to use simulation agents in this experiment because simulation agents do not select arguments based on compellingness. The result is provided in Table 45. The results in Table 45 support our hypothesis that a higher level noise in the historic data degrades the performance of inconsistency detection. We wonder whether this degradation would be brought into the detection of deception, but according to our experimental results in Table 46, detection performance is robust to the noise in historic data.

Table 45 Inconsistency detection performance with the level of noise in historic data

Inconsistency detection	Agents = 100, Evidence = 1-7, std. deviations # = 3, , Perturbation = real		
Noise in training data	*0.1*	*0.3*	*0.5*
Deceptive stories	0.203667	0.194333	0.171333
True stories	0.077333	0.069	0.059667

Table 46 Deception detection performance with the level of noise in historic data

Noise in training data	classification	precision	recall	F-measure	AUC
0.1	0.81875	0.784	0.88	0.829	0.858
0.3	0.81125	0.784	0.86	0.82	0.836
0.5	0.82875	0.804	0.87	0.836	0.857

A more realistic situation is that there might be deceptive stories in the historic data. We expect that as long as the majority of the historic data is honest, the performance of detection will remain satisfactorily high. To test this hypothesis, we simulated deceptive historic stories using deceptive reasoning in respectively 10%, 30% and 50% of the historic data. Table 47 presents the results of inconsistency detection, and Table 48 presents the results of deception detection. As more deceptive stories are contained in the historic data, the detection of inconsistency is less accurate. However, since deceptive stories are derived from natural reasoning, they do not introduce errors in other measures. Consequently, a moderate fraction of historic data being deceptive does not degrade the detection of deception.

Table 47 Inconsistency detection performance with the deception in historic data

Inconsistency detection	Agents = 100, Evidence = 1-7, std. deviations # = 3, Perturbation = real			
Deception in training data	*0%*	*10%*	*30%*	*50%*
Deceptive stories	0.203667	0.202333	0.199	0.194
True stories	0.077333	0.077333	0.074333	0.075

Table 48 Deception detection performance with the deception in historic data

Deception in training data	classification	precision	recall	F-measure	AUC
0%	0.81875	0.784	0.88	0.829	0.858
10%	0.8125	0.78	0.87	0.823	0.863
30%	0.815	0.789	0.86	0.823	0.871
50%	0.8175	0.793	0.86	0.825	0.873

4.7.3.3 Noise in Testing Data

Automatic argument retrieval introduces noise to the testing data. Unfortunately we cannot precisely evaluate the influence of automatic argument retrieval on inconsistency detection as we do not have the ground truth of inconsistency to the argument level. In this experiment, we mainly evaluate how strong the detection of deception depends on the noise introduced through argument retrieval. Below we compare the detection performance of manual argument retrieval (with small noise) and automatic argument retrieval (with large noise) using the abortion data. Table 49 presents the classification of true/deceptive stories, and Table 50 presents the classification of deceptive/non-deceptive stories. It seems that the classification rates and F-measures in both detections decrease. However, the detection using automatic argument retrieval still outperforms the verbal-cue methods and argument classification.

Table 49 Deception/truth classification performance with the argument retrievals

Abortion true/false	classification	precision	recall	F-measure	AUC
Manual argument retrieval	0.725	0.683	0.84	0.753	0.739
Automatic argument retrieval	0.71	0.684	0.78	0.729	0.781

Table 50 Deception/non-deception classification performance with the argument retrievals

Abortion misinformation	classification	precision	recall	F-measure	AUC
Manual argument retrieval	0.81875	0.784	0.88	0.829	0.858
Automatic argument retrieval	0.78125	0.761	0.82	0.789	0.867

4.7.3.4 Size of Evidence

During an inference, evidence imposes constraints on the reasoning results. We propose that the size of evidence has impact on the performance of deception detection, thus we first evaluate the impact of evidence on the inconsistency detection using simulated agents. Since deception only occurs in the testing process, our hypothesis is that the more evidence we provide the higher detection rate the system will achieve. The hypothesis can be explained intuitively by the fact that the more information we have about the

environment the more capable we are to detect any unusual observables. The results in Table 51 support our hypothesis.

Table 51 Detection performance with the size of evidence

Inconsistency detection	Agents = 10, Repeat =100, std. deviations # = 4, perturbation = 0.2						
Testing Evidence	1-5	6-10	11-15	16-20	21-25	26-30	31-35
True positive	0.9073	0.9403	0.9440	0.9576	0.9403	0.9447	0.9457
False positive	0.0440	0.0971	0.1507	0.2585	0.2673	0.3769	0.3639

From the observations in Chapter 4.6 that the detection rate of deception does not change significantly by replacing the evidence from the true datasets with the evidence from the deceptive datasets, it is reasonable to expect that this evidence does not have a strong impact on the detection of deception. To verify this expectation, we calculated the correlation between the size of evidence in a story and the deceptive score of the story from the classifier. The experiment shows that the correlation between the size of evidence and detectability is 0.078738, which is not significant. It agrees with our expectation that the size of evidence does not influence the performance of deception detection.

We now summarize the parametric study in terms of the parameters' impacts on inconsistency detection and deception detection. The significances of the impacts on inconsistency detection were also tested and verified by ANOVA. To evaluate the robustness of the model the parametric experiments on inconsistency detection were also performed on the other simulation networks including Hailfinder Network, Diabetes Network, and Munin Network. The result shows that although the detection rates vary from network to network, the influences of the parameters are the same. This, again,

proves that the method of inconsistency detection is robust to different structures and sizes of BNs as long as the network is ensured to have a moderate intra-dependency. A complete study that reports the parametric analysis on inconsistency detection can be found in Appendix B. The entire parametric study is summarized in Table 52.

Table 52 Summary of parametric study on inconsistency detection (ID) and deception detection (DD)

Parameter	*Group of agents*			*Individual agents*	
	group size	*similarity between agents*	*Deceivers #*	*inconsistency of agents*	*change in agent's story*
Significance to ID	Minor	Yes	Yes	Yes	No
Significance to DD	Expected No	No with condition	Expected Yes	Yes	No
Parameter	*Stories*				
	size of historic data	*noise in historic data*	*deception in historic data*	*noise in testing data*	*size of evidence*
Significance to ID	No	Yes	Yes	N/A	Yes
Significance to DD	Expected No	No	No	Yes	No

Among all 9 parameters, 6 parameters (similarity between agents, number of deceivers, inconsistency of agents, noise in historic data, deception in historic data and size of evidence) are found to significantly influence the performance of inconsistency detection, but deception detection is only sensitive to the number of deceivers, the inconsistency of agents, and the noise in testing data introduced by argument retrieval. The results can be explained as follows:

1. The size of the group determines the number of truth tellers who present opinions on the same topic as the deceiver. The more truth tellers we refer to, the better we can predict the opinions of the deceiver, and thus, it is easier to catch inconsistent

opinions. This impact is not explicit and is not expected to be brought into the detection of deception.

2. The similarity between agents determines how often the agents agree or disagree with each other. The deceiving agent's abnormal opinion will appear more distinct if the benevolent agents always agree or disagree with each other than if the benevolent agents have no clue about how the other agents will conclude. However, we do not find strong relationships between the coherence of the agents and the detection of deception except when an agent cannot be predicted by any other agent in the group.
3. The number of deceivers has a significant impact on the inconsistency detection. As soon as more than half of the agents becoming deceivers, the detection rate becomes unacceptably low. Together with the deceivers' influence on the distortion of expected truth we may be faced with unreliable deception detections.
4. The ground truth of inconsistency in the deceptive abortion data can be estimated by comparing the deceptive stories with the true stories. Combined with the results on the similarity between agents our method is proved to capture arguments that are internally inconsistent but not arguments that are different from others. By reducing inconsistency, deceivers can reduce the possibility of being detected.
5. However, being inconsistent does not indicate being changeable. Smaller differences between the testing story and the training story do not make agents less detectable.

6. The size of historic data determines how many repeats of data we can use to calculate the correlations between agents. Nevertheless, the number of repeats does not directly impact the performance of inconsistency detection, and thus is unlikely to impact the performance of deception detection either.
7. The noise in historic data influences the calculation of agent correlations, and hence reduces the detection rate of inconsistency. However, the degradation of the performance does not cause the decrease of deception detection rate.
8. When deceptive stories are mixed with true stories in historic data, the detection of inconsistency becomes inaccurate as it is harder to tell the normal behavior of a speaker. However, the errors do not interfere with other measures, nor cause the decrease of deception detection rate.
9. The noise in testing data can be due to the method used to retrieve semantics from natural-language stories such as the automatic argument retrieval using sentiment analysis. The noise can cause the decrease of deception detection rate, but even with noise in testing data our method still outperforms the verbal-cue methods.
10. The size of evidence determines how much knowledge we know before detecting inconsistency, thus more evidence can facilitate the detection of inconsistency. However, the size of evidence is not correlated with the detection of deception.

Chapter 5 Conclusion and Future Works

5.1 Overview

Deception and truth are two intertwined concepts since to tell a lie one needs to have some knowledge of the truth. Likewise, the understanding of deceivers is the key to telling the truth. A deceptive act is an intentional behavior meant to mislead the listener, and thus the intent of deception is a determinant of many behaviors of a deceiver. An effective detection method should enable the users to have a peek at the intent of the speaker. This is the insight that drives our research.

5.2 Contributions and Limitations

Inspired by many remarkable theories in human cognition and philosophical arguments on deception we learned that the reasoning process of deception can be regarded as normal reasoning with presupposing false conclusions. Thus, the first contribution of this dissertation is a proposal of a cognitive model of argumentation and deception. The model represents the knowledge of a speaker using a knowledge base and performs inference in a similar way to human reasoning.

The cognitive model of a speaker can be learned from his verbal content with the context knowledge supplied by similar speakers. Verbal content in the real world communications is uncertain, incomplete, and imprecise. The proposal of a learning method and its application on real test subjects are another key contribution of this work. With the extremely sparse and incomplete data from a subject's verbal content we are

still able to build a reasonable and realistic knowledge system that reflects the opinion of the subject, maximizes the uncertainty in his reasoning, and allows different possibilities of change in the opinion.

A third contribution, which is also the most important one, is the development of a detection model that detects deception by identifying deviations, explaining the deviations using the hypotheses of deceptive patterns, and combining the results for classification. This model is inspired by Johnson et al.'s (2001) process-based model and derived from the proposed cognitive model of deceivers. Driven by the intent, our model is able to discriminate unintentional misinformation such as misunderstanding, opinion change and wrong assumptions. Since the intent of deception cannot be avoided and is also hard to hide, our model detects deception in an effective and reliable fashion.

Other contributions lie with the analysis capability of the model to restore the truth, explain deceivers' tactics, provide evidence of deception and evaluate detection hypotheses computationally, which greatly facilitates the practical implementation of deception detection.

Limitations of this work arise from two aspects. The first aspect is with regards to the human reasoning process. As we have argued in this dissertation, the reasoning process of a deceiver is not guaranteed to deviate from a truth teller. Even backed up by different knowledge, a deceiver may still derive the same reasoning results as a truth teller. This could happen by chance or with competent deceivers who comprehensively master the

knowledge of a truth teller and carefully derive the arguments from the perspective of a truth teller. However even the best human or machine detectors have their limitations on the types of deception that they cannot detect, as there will likely to continue to exist deceptions that are undetectable. Another limitation of this work exists in the accuracy of semantics retrieval. Manual retrievals of the semantics are labor intensive and sometimes biased by the human raters' subjective interpretations. However, automatic sentiment retrieval techniques are still immature especially in understanding subjective information implied in the objective content. As our experiment shows, the detection of deception is sensitive to the error in the semantic arguments, therefore, the performance of this model is somewhat restrained by the accuracy of semantics retrieval.

5.3 Findings

The model of detection has been comprehensively evaluated using simulation data, synthesis data and real datasets. By comparing with the popular verbal-cue methods including bag of words, n-gram word features, POS tags and LIWC, we found our method to be the most accurate (effort in Chapter 4.4.2). Reliability issues that arise from the verbal-cue methods are that word features cannot be transferred from one topic to another and the discriminative words of a speaker cannot be applied to another, and that by re-wording, deceivers can significantly reduce the detection rate (effort in Chapter 4.4.2). We are not expected to have such problems because our model looks for universal patterns in deceptive reasoning that are independent of the domain knowledge and the wording of deceivers.

Two types of arguments were evaluated and compared: arguments based on facts and arguments based on opinions. We found that arguments based on opinions involve higher levels of reasoning, which is more complex than fact-based arguments (effort in Chapter 4.7.2.2), and that opinion-based arguments are less sensitive to evidence as people's subjective beliefs rarely change (effort in Chapter 4.6).

By studying different parameters of the datasets we found that the performance of detecting unexpected deviations is significantly influenced by the similarity between agents, the number of deceivers, the inconsistency of agents, the noise in historic data, the deception in historic data and the size of evidence (effort in Chapter 4.7). Not all of the parameters impact the detection of deception. Deception detection is only sensitive to the number of deceivers, the inconsistency of agents, and the noise in the testing data introduced by argument retrieval (effort in Chapter 4.7). These factors define the scope of deception where our method will be at a disadvantage. This scope contains the deception manipulated by multiple deceivers, the deception in which few manipulations are needed to distort the conclusion, and the deception in which arguments are not correctly represented.

5.4 Future Works

There is still room for improvement on the detection model, the selection of testbed, the evaluations, the sensitivity analysis and the post-analysis.

In terms of the detection model, extended studies can be carried out on the reasoning patterns. As we have implied in our expectations of the patterns with respect to different types of reasoning, the pattern "association of inconsistency and untruthfulness" seems to be most effective in discriminating deception and non-deception. We want to evaluate whether we are able to detect deception with this pattern alone. Currently we assume independence between the patterns. However, interactions between the patterns may exist. By incorporating the interactions between the patterns in the detection model, the performance may be further improved. We are also in search of new patterns. More effective patterns can be derived by borrowing ideas from the cueing methods. Another improvement regards the semantic arguments retrieved from the verbal content. Although people's attitudes of an argument are usually expressed as polarities, their degrees of belief are embedded in the phrases, sometimes with certainty modifiers and sometimes with arguments on both sides. The degree of belief is lost if only the polarity of arguments is encoded. Actually, existing sentiment analysis tools can automatically retrieve the degree of a polarity instead of the binary states of arguments. We want to evaluate the performance of the tools and their impact on deception detection.

During the evaluation, a pilot study was conducted to evaluate the situation with multiple deceivers. A natural extension on this testbed is to consider the interaction between the agents in a group. According to (Hung, 2012), deceivers who collaborate to dupe others behave differently while their "partners in crime" are speaking. In addition, truth tellers may also be influenced by deceptive speakers as a result of social pressure (Asch, 1955).

In particular, we want to embed the interaction between deceivers and the interaction between deceivers and truth tellers into the testbed.

In terms of the evaluation we expect that our model is more reliable than verbal-cue methods because the retrieval of arguments is insensitive to the wording and the scoring of the deceptive patterns are independent of the topic. Empirical experiments can be performed to verify this expectation. In particular, we can use automatic sentiment analysis tools to extract the polarities of arguments based on the stories in which the discriminative words are replaced by synonyms. The classifier learned from one dataset can be used to classify stories of another dataset in order to evaluate the patterns' dependence on topic; and the classifier learned from a part of a dataset can be used to classify the other part of the dataset in order to evaluate the patterns' dependence on individual speakers.

During the construction of the cognitive models we have made a few assumptions such as that the arguments can be accurately extracted, the speakers are fairly uncertain about the arguments, and ignored arguments are close to what the majority agrees. We have evaluated the sensitivity of a speaker's cognitive model to these assumptions, but according to the analysis in Chapter 4.7.3.3, which evaluates the performance of detection with regards to the retrieval of arguments, we observed that errors introduced to the construction of cognitive models can be brought into deception detection. A thorough analysis can be performed by detecting deception using erroneous cognitive models

caused by automatic methods of selecting topics, higher levels of uncertainty in the cognitive models, lower weights of an individual's story in the learning data, and etc.

In terms of post-analysis, another important task is to study deceivers' behavior. We have argued that some deceivers are undetectable through investigating their reasoning results. An interesting problem would be what kind of deceiver is less detectable or more credible. Bond and DePaulo have in their paper (Bond and Depaulo, 2008) suggested that some deceivers are more credible than others whether lying or truth-telling, and some speakers are more detectable meaning that the validity of their statements is easier to be detected. We believe that the credibility and the detectability of speakers are determined by their knowledge structures. A future direction would be to investigate what structure enables a deceiver to infer more convincing results and what structure makes an agent significantly truthful when telling truth and significantly untruthful when telling lies. This can be helpful in quantifying and explaining verbal cues and shed lights on the discovery of more effective verbal cues.

Appendix

Appendix A- Experimental results of Intra-dependency

According to Yuan (2007), detection rate is largely influenced by the network's intra-dependency. The intra-dependency index measures how dependent the states' probabilities are on the evidence. It can be calculated using (Yuan, 2007):

$$I = \frac{\sum_{j=1}^{M} \sqrt{\sum_{i=1}^{N} (R_{i,j} - \overline{R}_i)^2}}{NM}$$

where $R_{i,j}$ denotes the posterior probability of r.v. i in the jth test, $\overline{R}_i$ is the "neutral" value of r.v. i, N is the number of variables in the network, and M is the total number of tests. The "neutral" value of a r.v. is the average of all probabilities that the variable has obtained over all the test cases, which is calculated using:

$$\overline{R}_i = \frac{\sum_{j=1}^{M} R_{i,j}}{M}$$

Normally, the farther away a node is from the evidence, the less strongly it depends on the evidence. Since the nodes in Diabetes network are highly separated from one another due to its larger height, we form the hypothesis that the nodes' dependency on the evidence is the weakest among all networks we tested on. To confirm our hypothesis, an experiment was conducted to measure the intra-dependency indices of all the networks. Table 53 shows the test result. The result confirms our hypothesis that Diabetes Network has the lowest intra-dependency.

Table 53 Intra-dependency index of different networks

Network	Intra-Dependency Index
Alarm	0.023946267072262
Hailfinder	0.011272353508164927
Diabetes	0.001618689030060108
Munin	0.002419162273260489

In Yuan (2007), parameters that influence the intra-dependency index were also studied. It demonstrates that the amount of evidence and the range of perturbation used in the multi-agent experiments mainly determine the intra-dependency of the nodes. This is due to the fact that the more evidence we possess, the more strongly the nodes depend on the evidence, but the dependency turns out to be weaker if the agents are perturbed more heavily. In addition to these two parameters, we showed that the structure of the network, specifically the height, also impacts the intra-dependency.

Appendix B- Parametric experimental results of identification of inconsistency

An exhaustive definite study of this can be found in (Santos and Li, 2010). The results presented here are updated due to a bug in the program.

1) Results on the number of agents and the perturbation value:
Table 54(a) shows the means of Pearson correlation values of all states. As we can see, the Pearson correlation values are only determined by perturbation values. This is because the more heavily we perturb the agents, the less correlated the agents are. Table 54(b) shows the means of the standard deviations of the prediction error. It seems that the standard deviation has a slightly negative correlation with the number of agents. This can be explained by the fact that having more agents increases the number of correlation values for each agent, and thus increases the precision of predicting opinions. On the contrary, the perturbation value has a significant influence on the standard deviation because the less correlated the agents are, the more difficult it is to predict their opinions. Overall, the result shows that the more correlated the agents are, the more obvious the inconsistency appears to be.

Table 54 Detection performance with the number of agents. (a) means of Pearson correlation values. (b) means of prediction error stdev. (c) means of positive detection rate. (d) means of false detection rate.

(a)				
Parameters	Repeats = 100, Evidence = 10, No. of stdevs = 4			
Pert.\ Agents	3	10	30	100
0.1	0.904056	0.914231	0.902494	0.893539
0.2	0.829023	0.82672	0.826863	0.817745
0.3	0.729068	0.755321	0.750651	0.761878
0.4	0.701626	0.707899	0.698671	0.699698
(b)				
Parameters	Repeats = 100, Evidence = 10, No. of stdevs = 4			
Pert.\ Agents	3	10	30	100
0.1	0.0434926	0.0362296	0.0375804	0.0371846
0.2	0.0468883	0.0424275	0.0390635	0.0406773
0.3	0.0490258	0.0424262	0.0402245	0.0390285
0.4	0.0558387	0.0439094	0.0408072	0.0399069
(c)				
Parameters	Repeats = 100, Evidence = 10, No. of stdevs = 4			
Pert.\ Agents	3	10	30	100
0.1	0.901392	0.940865	0.923081	0.932075
0.2	0.838475	0.848329	0.86418	0.856521
0.3	0.770829	0.791666	0.799054	0.810289
0.4	0.606502	0.715911	0.744098	0.759478
(d)				
Parameters	Repeats = 100, Evidence = 10, No. of stdevs = 4			
Pert.\ Agents	3	10	30	100
0.1	0.0192964	0.0201131	0.0162917	0.0143921
0.2	0.0163721	0.0142754	0.0154445	0.0157893
0.3	0.0186278	0.0178021	0.0224165	0.0213662
0.4	0.0098979	0.0172704	0.019955	0.0223486

2) Results on the number of repeats:

The result in Table 55 does not show significant correlation between repeats and detection.

Table 55 Detection performance with the number of repeats. (a) means of Pearson correlation values. (b) means of prediction error stdev. (c) means of positive detection rate. (d) means of false detection rate.

(a)				
Parameters	Agents=10, Evidence=10, No. of stdevs=4			
Pert.\ Repeats	*10*	*100*	*1000*	*10000*
0.1	0.895688	0.914231	0.907112	0.902621
0.2	0.786274	0.82672	0.823928	0.825596
0.3	0.711214	0.755321	0.756259	0.765995
0.4	0.695493	0.707899	0.69208	0.692727
(b)				
Parameters	Agents=10, Evidence=10, No. of stdevs=4			
Pert.\ Repeats	*10*	*100*	*1000*	*10000*
0.1	0.0352783	0.0362296	0.0387869	0.0388422
0.2	0.0354494	0.0424275	0.0434013	0.0429281
0.3	0.0389239	0.0424262	0.0444829	0.0445376
0.4	0.0393172	0.0439094	0.0421055	0.0440204
(c)				
Parameters	Agents=10, Evidence=10, No. of stdevs=4			
Pert.\ Repeats	*10*	*100*	*1000*	*10000*
0.1	0.914639	0.940865	0.930501	0.928079
0.2	0.887225	0.848329	0.854309	0.860704
0.3	0.849395	0.791666	0.782353	0.776183
0.4	0.792878	0.715911	0.727012	0.722417
(d)				
Parameters	Agents=10, Evidence=10, No. of stdevs=4			
Pert.\ Repeats	*10*	*100*	*1000*	*10000*
0.1	0.0617718	0.0201131	0.0121549	0.0132868
0.2	0.109151	0.0142754	0.0107823	0.0111045
0.3	0.0868499	0.0178021	0.0111043	0.0106883
0.4	0.0645721	0.0172704	0.0138269	0.0105358

3) Results on the amount of evidence in the testing process:

The hypothesis can be explained intuitively by the fact that the more information we have about the environment, the easier for us to identify any abnormal phenomenon. The results in Table 56 support our hypothesis.

Table 56 Detection performance with the number of pieces of evidence in the testing process. (a) means of Pearson correlation values. (b) means of prediction error stdev. (c) means of positive detection rate. (d) means of false detection rate

(a)							
Parameters	Repeats = 100, Agents = 10, Training Evidence = 1-5, No. of stdevs = 4						
Pert.\ Test. evi.	*1-5*	*6-10*	*11-15*	*16-20*	*21-25*	*26-30*	*31-35*
0.1	0.906734	0.904841	0.908053	0.904443	0.906235	0.905514	0.905574
0.2	0.826037	0.818982	0.821374	0.817364	0.818707	0.806348	0.826104
0.3	0.767322	0.761859	0.749494	0.762457	0.743148	0.742437	0.753039
0.4	0.6743	0.668007	0.697072	0.699263	0.680636	0.692426	0.701311
(b)							
Parameters	Repeats = 100, Agents = 10, Training Evidence = 1-5, No. of stdevs = 4						
Pert.\ Test. evi.	*1-5*	*6-10*	*11-15*	*16-20*	*21-25*	*26-30*	*31-35*
0.1	0.0264153	0.0264478	0.028317	0.0278283	0.0277945	0.027646	0.0294502
0.2	0.0302722	0.0288494	0.0304743	0.0281466	0.0321399	0.028686	0.0298523
0.3	0.0325099	0.0337676	0.0305145	0.0314473	0.0307659	0.0317314	0.0328154
0.4	0.0330761	0.0313959	0.0321072	0.0293773	0.0310706	0.0314706	0.0340528
(c)							
Parameters	Repeats = 100, Agents = 10, Training Evidence = 1-5, No. of stdevs = 4						
Pert.\ Test. evi.	*1-5*	*6-10*	*11-15*	*16-20*	*21-25*	*26-30*	*31-35*
0.1	0.96059	0.966432	0.968372	0.96055	0.96131	0.952192	0.952818
0.2	0.907332	0.940319	0.944091	0.957625	0.940315	0.944709	0.945776
0.3	0.865615	0.896799	0.922156	0.922417	0.940554	0.941941	0.933806
0.4	0.820115	0.848268	0.873377	0.917588	0.917499	0.915642	0.925681
(d)							
Parameters	Repeats = 100, Agents = 10, Training Evidence = 1-5, No. of stdevs = 4						
Pert.\ Test. evi.	*1-5*	*6-10*	*11-15*	*16-20*	*21-25*	*26-30*	*31-35*
0.1	0.0263437	0.0762742	0.145813	0.19918	0.241546	0.289513	0.269343
0.2	0.0440153	0.0971125	0.150783	0.258578	0.267313	0.376949	0.363976
0.3	0.0408427	0.0683694	0.17988	0.248691	0.345157	0.390614	0.441941
0.4	0.0485377	0.095276	0.181732	0.297085	0.366939	0.415675	0.485484

4) Results on the amount of evidence in the training process:

The number of training evidence indicates the size of evidence space. 75% of evidence has the largest evidence space, which results in the training data that is irrelevant to the testing data. The experimental results are shown in Table 57.

Table 57 Detection performance with the number of pieces of evidence in the training process. (a) means of Pearson correlation values. (b) means of prediction error stdev. (c) means of positive detection rate. (d) means of false detection rate

(a)							
Parameters	Repeats = 100, Agents = 10, Testing Evidence = 1-5, No. of stdevs = 4						
Pert.\ Train. evi.	*1-5*	*6-10*	*11-15*	*16-20*	*21-25*	*26-30*	*31-35*
0.1	0.89816	0.913618	0.922451	0.904141	0.906847	0.917435	0.887753
0.2	0.823716	0.834601	0.854739	0.841038	0.833634	0.849105	0.834878
0.3	0.737116	0.775518	0.779752	0.789052	0.79364	0.797002	0.796899
0.4	0.703018	0.699213	0.724156	0.747219	0.745863	0.755436	0.729103
(b)							
Parameters	Repeats = 100, Agents = 10, Testing Evidence = 1-5, No. of stdevs = 4						
Pert.\ Train. evi.	*1-5*	*6-10*	*11-15*	*16-20*	*21-25*	*26-30*	*31-35*
0.1	0.0260764	0.0429822	0.0557096	0.0681927	0.0767133	0.0779643	0.0775439
0.2	0.029799	0.0521842	0.0655847	0.0813936	0.0895538	0.0994152	0.0958367
0.3	0.0304199	0.0548619	0.0707517	0.0847784	0.0936349	0.101635	0.103183
0.4	0.0333709	0.0561343	0.0725266	0.0891179	0.0984678	0.105822	0.107264
(c)							
Parameters	Repeats = 100, Agents = 10, Testing Evidence = 1-5, No. of stdevs = 4						
Pert.\ Train. evi.	*1-5*	*6-10*	*11-15*	*16-20*	*21-25*	*26-30*	*31-35*
0.1	0.959565	0.893376	0.860812	0.799447	0.782748	0.780562	0.797792
0.2	0.897879	0.768845	0.685397	0.589983	0.533972	0.480152	0.526503
0.3	0.825663	0.655597	0.548082	0.434086	0.403563	0.352897	0.396509
0.4	0.734696	0.577444	0.48275	0.366147	0.305052	0.292564	0.322489
(d)							
Parameters	Repeats = 100, Agents = 10, Testing Evidence = 1-5, No. of stdevs = 4						
Pert.\ Train. evi.	*1-5*	*6-10*	*11-15*	*16-20*	*21-25*	*26-30*	*31-35*
0.1	0.0314239	0.0023884	0.0008331	0.0006186	0.0012465	0.0057588	0.0521203
0.2	0.0183397	0.0012748	0.0004945	0.0002918	0.0025096	0.0006032	0.0097322
0.3	0.0245182	0.0019625	0.0005257	0.0002063	2.875E-52	0.000792	0.0217482
0.4	0.0238318	0.0034444	5.66E-57	0.0001156	2.938E-52	0.0002653	0.0119077

5) Results on the number of standard deviations:

The results shown in Table 58 indicate that if we relax the number of standard deviations, we will get fewer positive and negative alarms. This is very intuitive to understand since the more forgiving we are, the fewer inconsistencies we will care about.

Table 58 Detection performance with the number of standard deviation. (a) means of Pearson correlation values. (b) means of prediction error stdev. (c) means of positive detection rate. (d) means of false detection rate

(a)				
Parameters	Agents=10, Repeats=100, Evidence=10			
Pert.\ No. of stdevs	*4*	*3*	*2*	*1*
0.1	0.914231	0.905259	0.894034	0.902577
0.2	0.82672	0.818914	0.821617	0.822027
0.3	0.755321	0.759824	0.767907	0.760969
0.4	0.707899	0.696742	0.702926	0.705274
(b)				
Parameters	Agents=10, Repeats=100, Evidence=10			
Pert.\ No. of stdevs	*4*	*3*	*2*	*1*
0.1	0.0362296	0.0362954	0.0345131	0.0373867
0.2	0.0424275	0.0388624	0.0429311	0.0426748
0.3	0.0424262	0.0399345	0.0427087	0.0432754
0.4	0.0439094	0.0413434	0.0416888	0.0457272
(c)				
Parameters	Agents=10, Repeats=100, Evidence=10			
Pert.\ No. of stdevs	*4*	*3*	*2*	*1*
0.1	0.940865	0.968434	0.983273	0.995557
0.2	0.848329	0.912573	0.966484	0.988483
0.3	0.791666	0.884303	0.934539	0.980235
0.4	0.715911	0.836423	0.919155	0.965634
(d)				
Parameters	Agents=10, Repeats=100, Evidence=10			
Pert\ No. of stdevs	*4*	*3*	*2*	*1*
0.1	0.0201131	0.039236	0.134116	0.316018
0.2	0.0142754	0.0521267	0.106949	0.296226
0.3	0.0178021	0.0524848	0.127107	0.290299
0.4	0.0172704	0.0548502	0.121914	0.321566

From Table 54 to Table 58, we can also see that except when the training evidence is not relevant to the testing evidence, the detection rate is always above 60% (higher than human detection rate). Therefore, to ensure a good detection performance which is robust to environmental change, it is necessary to make sure that the communication environment for the history data is consistent with the environment for the current data.

Appendix C- Learning of the Reasoning Process

The structure of the knowledge base is built based on the following steps.

Step 1: Retrieval of the arguments
We decided to use 30 arguments to represent the nodes in the BN since 30 nodes form a BN with moderate size and 30 topics contain the majority of the arguments presented in both true and false stories. Firstly, we remove the punctuations of the stories, which includes ,, ---, --, *:* , *;* , *"*, *(,)* , *!*, */*, *?* and .. Secondly, we use an open-source program called "snowball" to stem all the words into their root forms. Then we build a dictionary of all words and a vector for each story to indicate the frequency of each word that exists in the story. After the dictionary is built, we retrieve the most frequent *N* words over all stories, from which we manually summarize 30 topics. *N* is currently set as a number slightly larger than 100. It may vary according to the dataset.

Step 2: Retrieval of the sentiments
We use key word mapping to map a sentence with one or several arguments. Specifically, we select some key words to represent an argument, e.g. *right*, *choose*, *control*, *decide*, *own*, *body*, *woman*, *mother*, *every* and their variations represent the argument that *Women have the right to do whatever they want with their bodies*. If a sentence contains one or several of the key words, the story which contains the sentence is said to contain the argument. The polarity of the sentence with regards to this argument is also retrieved by mapping key words that represent negativity such as *cannot*, *never*, *isn't*. Similar techniques have been used in sentiment analysis (Liu, 2010). The sentiments are filled into a matrix with *m* rows and *n* columns where *m* represents the number of stories and *n* represents the number of arguments, and are encoded into 1 if the positive attitude of the argument is expressed and 0 if the negative attitude is expressed. After the automatic generation of the sentiments, we carefully go through all the stories to confirm the results. We eliminate entries that do not have the semantic meaning of the mapped arguments and change the polarities that were mapped wrong.

Step 3: Building BN structures
We use PC algorithm to build the structure of a BN. The basic idea of PC algorithm is that partial correlation between r.v.s indicates d-separation. PC algorithm can be divided into two parts: connection and orientation. Below is a pseudo code of PC algorithm.

Let V denote the set of all nodes. Let ADJ_x denote the set of nodes that are adjacent to node x. Let $p(x,y|S)$ denote the partial correlation between x and y given set S.
$I(x,y|S)$ represents the state that x and y are independent given S. $I(x,y|S)$=true if $p(x,y|S)$<*some threshold*

1. Start with a complete undirected graph g
2. $i = 0$
3. While i<some threshold
   - 3.1 For each node $x \in V$
     - 3.1.1 For each $y \in ADJ_x$
       - 3.1.1.1 Determine if there is $S \subseteq ADJ_x - y$ with $|S| = i$ and $I(x,y|S)$
       - 3.1.1.2 If this set exists, remove link between x and y from g
     - 3.1.2 Until $|ADJ_x| < i$
   - 3.2 $i = i+1$
4. For each uncoupled structure x-z-y
   - 4.1 If $z \in S_{x,y}$, Orient x-z-y as $x \rightarrow z \leftarrow y$
5. While no more edges can be oriented from 4
   - 5.1 For each uncoupled structure $x \rightarrow z - y$
     - 5.1.1 Orient $z - y$ as $z \rightarrow y$
   - 5.2 For each x-z such that there is a directed path from x to z
     - 5.2.1 Orient $x - z$ as $x \rightarrow z$
   - 5.3 For each uncoupled structure $x - z - y$ such that $x \rightarrow w$, $y \rightarrow w$ and $z - w$
     - 5.3.1 Orient $z - w$ as $z \rightarrow w$

Appendix D- Conditional Probabilities of the Bayesian Network of A's Belief System

B_relation_with_A_s_mother

good	bad
0.2	0.8

A_hate_Indian

A_have_exp_of_prostitution	B_relation_with_A_s_mother	T	F
T	good	0.1	0.9
	bad	0.3	0.7
F	good	0.5	0.5
	bad	0.55	0.45

B_relation_with_A

A_is_nice_to_B	B_relation_with_A_s_mother	rape	fan
T	good	0.4	0.6
	bad	0.4	0.6
F	good	0.5	0.5
	bad	0.8	0.2

B_is_new_to_party

T	F
0.8	0.2

A_is_nice_to_B

A_hate_Indian	T	F
T	0.3	0.7
F	0.7	0.3

A_have_drug_from

B_is_new_to_party	B_relation_with_A	B	self
T	rape	0.7	0.3
	fan	0.1	0.9
F	rape	0.9	0.1
	fan	0.2	0.8

B_knows_A_s_adr

B_relation_with_A	T	F
rape	0.7	0.3
fan	0.7	0.3

B_in_A_s_party_by

A_hate_Indian	B_relation_with_A	self	invitation
T	rape	0.9	0.1
	fan	0.6	0.4
F	rape	0.5	0.5
	fan	0.4	0.6

B_drive_A_home

B_knows_A_s_adr	B_in_A_s_party_by	T	F
T	self	0.9	0.1
	invitation	0.9	0.1
F	self	0.1	0.9
	invitation	0.1	0.9

A_have_exp_of_prostitution

T	F
0.8	0.2

cry_for_help

sex_by	T	F
rape	0.8	0.2
enticement	0.2	0.8

A_is_celebrity

T	F
0.8	0.2

A_s_boyfriend__catch_on_the_scene

T	F
0.8	0.2

B_refuse_to_pay

T	F
0.8	0.2

A_claim_being_raped

B_refuse_to_pay	A_is_celebrity	sex_by	A_s_boyfriend__catch_on_the_scene	T	F
T	T	rape	T	0.6	0.4
			F	0.6	0.4
		enticement	T	0.7	0.3
			F	0.6	0.4
	F	rape	T	0.8	0.2
			F	0.8	0.2
		enticement	T	0.9	0.1
			F	0.7	0.3
F	T	rape	T	0.6	0.4
			F	0.6	0.4
		enticement	T	0.5	0.5
			F	0.1	0.9
	F	rape	T	0.8	0.2
			F	0.8	0.2
		enticement	T	0.8	0.2
			F	0.2	0.8

sex_by

A_have_drug_from	B_drive_A_home	A_have_exp_of_prostitution	A_hate_Indian	rape	enticement
B	T	T	T	0.8	0.2
			F	0.5	0.5
		F	T	0.9	0.1
			F	0.6	0.4
	F	T	T	0.5	0.5
			F	0.5	0.5
		F	T	0.5	0.5
			F	0.5	0.5
self	T	T	T	0.6	0.4
			F	0.3	0.7
		F	T	0.7	0.3
			F	0.4	0.6
	F	T	T	0.5	0.5
			F	0.5	0.5
		F	T	0.5	0.5
			F	0.5	0.5

Appendix E- Sample data from the abortion dataset and the hotel reviews

Abortion data (Mihalcea et al., 2009):
Abortions should be legal. I think that women have the right to choose what to do with their own bodies. A government should not interfere with how a woman deals with her pregnancy. Only she knows the truth behind the conception--- there could be birth defects, maybe she was raped or had an incestuous relationship. She should be able to decide if she wants to terminate her pregnancy.

Hotel review (Ott et al., 2011):
We stay at Hilton for 4 nights last march. It was a pleasant stay. We got a large room with 2 double beds and 2 bathrooms, The TV was Ok, a 27' CRT Flat Screen. The coincierge was very friendly when we need. The room was very cleaned when we arrived, we ordered some pizzas from room service and the pizza was Ok also.The main Hall is beautiful. The breakfast is charged, 20 dollars, kinda expensive. The internet access (WiFi) is charged, 13 dollars/day. Pros: Low rate price, huge rooms, close to attractions at Loop, close to metro station. Cons: Expensive breakfast, Internet access charged. Tip: When leaving the building, always use the Michigan Av exit. Its a great view.

Appendix F- A tutorial on Bayesian Networks

A Bayesian Network is an annotated directed acyclic graph (DAG), which is composed of nodes and arcs. Nodes store knowledge in the form of random variables, and directed arcs connecting two nodes represent a conditional/causal relationship between them. The uncertainty of the relationship is encoded in a conditional probability. The conditional probabilities between any random variable and its parents are contained in an associated conditional probability table (CPT). Under the conditional independence assumption, the chain rule, which is also the product of the CPTs, is expressed as

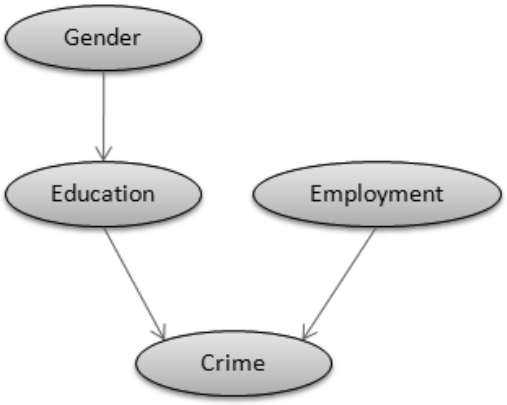
$$P(X_1, X_2, ..., X_n) = \prod_1^n P(X_i \mid parents(X_i))$$

This provides a representation of the joint probability distribution, with which a BN is able to present the direct relationships between variables and form a structural organization of information.

Figure 13 is a simple example of a BN. It represents the relationship between possible causes and consequences of committing a crime. Each random variable in the example has two states. The arcs between each two nodes denote the causal relationship between possible states of the two random variables. For example, if someone is a male, then his education level is above high school with a probability of 0.65. The roots of the network (Gender and Employment in this case) have prior probabilities instead of conditional probabilities, which represent the probability of a person being male and that of a person being employed regardless of any evidence.

Reasoning in BKBs comes in two types: belief updating and belief revision. Bayesian updating calculates the posterior probability of each node given some evidence. Belief revision identifies the most probable instantiation of all random variables with given evidence by computing joint probability $P(A_{k+1}=a_{k+1}, \ldots, A_n=a_n | A_1=a_1, \ldots, A_k=a_k)$ where A_i denotes the i^{th} node, and $A_1=a_1, \ldots, A_k=a_k$ are evidence through applying the chain rule. The most probable instantiation of all random variables is also called the most probable explanation (MPE). Taking Figure 13 as an example, after comparing the joint probabilities of the two inferences, we find that the instatiations of the nodes Gender=male, Education=above high school, Employment=yes and Crime=no have the highest joint probability:

$$\begin{aligned}
&\mathrm{P}(\mathrm{Gender} = \mathrm{male}) \\
&\times \mathrm{P}(\mathrm{Education} = \mathrm{above\ high\ school} | \mathrm{Gender} = \mathrm{male}) \\
&\times \mathrm{P}(\mathrm{Employment} = \mathrm{yes}) \\
&\times \mathrm{P}(\mathrm{Crime} = \mathrm{no} | \mathrm{Education} = \mathrm{above\ high\ school}, \mathrm{Employment} = \mathrm{yes}) \\
&= 0.6 \times 0.65 \times 0.93 \times 0.99 = 0.359073 .
\end{aligned}$$



P(Gender=male) = 0.6
P(Gender=female) = 0.4
P(Education=above high school | Gender=male) = 0.65
P(Education=above high school | Gender=female) = 0.73
P(Employment=yes) = 0.93
P(Employment=no) = 0.07
P(Crime=yes | Education=above high school,
Employment=yes) = 0.01
P(Crime=yes | Education=above high school,
Employment=no) = 0.05
P(Crime=yes | Education=below high school,
Employment=yes) = 0.1
P(Crime=yes | Education=below high school,
Employment=no) = 0.2

Figure 13 A simple BN example

References

1. Ford, Charles V., and John S. Price. *Lies!, lies!!, lies!!!: The psychology of deceit*. Washington, DC: American Psychiatric Press, 1996.

2. Millar, Murray G., and Karen U. Millar. "The effects of cognitive capacity and suspicion on truth bias." *Communication Research* 24.5 (1997): 556-570.

3. Bond, Charles F., and Bella M. DePaulo. "Individual differences in judging deception: Accuracy and bias." *Psychological bulletin* 134.4 (2008): 477.

4. Johnson, Paul E., et al. "Detecting deception: adversarial problem solving in a low base-rate world." *Cognitive Science* 25.3 (2001): 355-392.

5. George, Joey F., and John R. Carlson. "Group support systems and deceptive communication." *System Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on*. IEEE, 1999.

6. Zhou, Lina, and Zhang, Dongsong, "Automatic deception detection in computer-mediated communication", *Intelligent Systems, IEEE* 27.6 (2012): 73-75.

7. Burgoon, Judee K., and David B. Buller. "Interpersonal deception: III. Effects of deceit on perceived communication and nonverbal behavior dynamics." *Journal of Nonverbal Behavior* 18.2 (1994): 155-184.

8. McCornack, Steven A., and Timothy R. Levine. "When lovers become leery: The relationship between suspicion and accuracy in detecting deception."*Communications Monographs* 57.3 (1990): 219-230.

9. Zajonc, Robert B., Alexander Heingartner, and Edward M. Herman. "Social enhancement and impairment of performance in the cockroach." *Journal of Personality and Social Psychology* 13.2 (1969): 83.

10. George, Joey F., and Kent Marett. "Inhibiting deception and its detection."*System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*. IEEE, 2004.

11. Inbau, Fred Edward. *Lie detection and criminal interrogation*. Williams & Wilkins Co., 1948.

12. Hayley, Hung, "Deception Detection in Multiparty Contexts", *Intelligent Systems*, *IEEE* 27.6 (2012): 69-71.

13. DePaulo, Bella M., et al. "Cues to deception." *Psychological bulletin* 129.1 (2003): 74.

14. Cybenko, George, Annarita Giani, and Paul Thompson. "Cognitive hacking: A battle for the mind." *Computer* 35.8 (2002): 50-56.

15. Miller, David, ed. *Tell me lies: Propaganda and media distortion in the attack on Iraq*. London: Pluto, 2004.

16. Santos, Eugene, Jr., et al. "Intelligence Analyses and the Insider Threat." *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*42.2 (2012): 331-347.

17. Weintraub, Jerry (producer) & Soderbergh, Steven (director). Ocean's Eleven [Motion picture]. United States: Warner Bros.

18. Whaley, Barton. "Toward a general theory of deception." *The Journal of Strategic Studies* 5.1 (1982): 178-192.

19. Baron, Marcia. "What is wrong with self-deception." *Perspectives on self-deception* (1988): 431-449.

20. Barnes, Annette. *Seeing through self-deception*. Cambridge University Press, 2007.

21. Mahon, James Edwin. "A definition of deceiving." *International Journal of Applied Philosophy* 21.2 (2008): 181-194.

22. Demos, Raphael. "Lying to oneself." *The Journal of Philosophy* 57.18 (1960): 588-595.

23. Chisholm, Roderick M., and Thomas D. Feehan. "The intent to deceive." *The Journal of Philosophy* 74.3 (1977): 143-159.

24. Adler, Jonathan E. "Lying, deceiving, or falsely implicating." *The Journal of Philosophy* 94.9 (1997): 435-452.

25. Gert, Bernard. *Morality: Its nature and justification*. Oxford University Press on Demand, 1998.

26. Fuller, Gary. "Other-Deception." *Southwestern Journal of Philosophy* 7.1 (2008): 21-31.

27. Linsky, Leonard. "Deception." *Inquiry* 6.1-4 (1963): 157-169.

28. Van Horne, Winston A. "Prolegomenon to a Theory of Deception." *Philosophy and Phenomenological Research* 42.2 (1981): 171-182.

29. Bok, Sissela. *Lying: Moral choice in public and private life*. Vintage, 2011.

30. Barnes, John Arundel. *A pack of lies: Towards a sociology of lying*. Cambridge University Press, 1994.

31. Davidson, Donald. "Deception and division." *The multiple self* (1987): 79.

32. Ryle, Gilbert. *The concept of mind*. University of Chicago Press, 1949.

33. Schmitt, Frederick F. "Epistemic dimensions of self-deception." *Perspectives on self-deception* 6 (1988): 183.

34. Vrij, Aldert. *Detecting lies and deceit: Pitfalls and opportunities*. Wiley-Interscience, 2008.

35. O'Neill, Barry. "A formal system for understanding lies and deceit." *Jerusalem Conference on Biblical Economics*. 2003.

36. Ekman, Paul. *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. WW Norton & Company, 2009.

37. Scott, Gini Graham. *The truth about lying*. Booktango, 1994.

38. Pinto, Robert C. *Argument, inference and dialectic: Collected papers on informal logic*. Vol. 4. Kluwer Academic Pub, 2001.

39. Levi, Isaac. *For the Sake of the Argument: Ramsey Test Conditionals, Inductive Inference, and Nonmonolitic Reasoning*. Cambridge University Press, 1996.

40. Goldman, Alvin I. *A theory of human action*. Vol. 2. Englewood Cliffs, NJ: Prentice-Hall, 1970.

41. Davidson, Donald. "Mental events." *Readings in philosophy of psychology* 1 (1980): 107-119.

42. Mele, Alfred R. *Springs of action: Understanding intentional behavior*. Oxford University Press on Demand, 1992.

43. Humberstone, I. Lloyd. "Direction of fit." *Mind* 101.401 (1992): 59-83.

44. Velleman, J. David. "The guise of the good." *Nous* 26.1 (1992): 3-26.

45. Vasilyeva, Alena, and Mark Frank. "Testing Interpersonal Deception Theory: Strategic and Nonstrategic Behaviors of Deceivers and Truth Tellers, Communication Skills, and Dynamic Character of Deception". The annual meeting of the International Communication Association, Dresden International Congress Centre, Dresden, Germany , March, 2011.

46. Bell, J. Bowyer, and Barton Whaley. *Cheating and deception*. Transaction Publishers, 1991.

47. Wile, Ira S. "Lying as a biological and social phenomenon." *Nervous Child* 1 (1942): 293-313.

48. Sodian, Beate. "The development of deception in young children." *British Journal of Developmental Psychology* 9.1 (1991): 173-188.

49. Cruickshank, Charles Greig, and C. Cruickshank. *Deception in World War II*. Oxford: Oxford University Press, 1979.

50. Brault, Sébastien, et al. "Deception in Sports Using Immersive Environments." *Intelligent Systems, IEEE* 27.6 (2012): 64-66.

51. Grazioli, Stefano, and Sirkka L. Jarvenpaa. "Perils of Internet fraud: An empirical investigation of deception and trust with experienced Internet consumers."*Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 30.4 (2000): 395-410.

52. Lieberman, David J. *Never be lied to again: How to get the truth in 5 minutes or less in any conversation or situation*. St. Martin's Griffin, 1999.

53. Mihalcea, Rada, and Carlo Strapparava. "The lie detector: Explorations in the automatic recognition of deceptive language." *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, 2009.

54. DePaulo, Bella M., Julie I. Stone, and G. Daniel Lassiter. "Deceiving and detecting deceit." *The self and social life* 323 (1985).

55. Pennebaker, James W., et al. "The development and psychometric properties of LIWC2007." *Austin, TX, LIWC. Net* (2007).

56. Rubin, Victoria L., and Tatiana Vashchilko. "Identification of truth and deception in text: Application of vector space model to rhetorical structure theory."*Proceedings of the Workshop on Computational Approaches to Deception Detection*. Association for Computational Linguistics, 2012.

57. Ott, Myle, et al. "Finding deceptive opinion spam by any stretch of the imagination." *arXiv preprint arXiv:1107.4557* (2011).

58. Newman, Matthew L., et al. "Lying words: Predicting deception from linguistic styles." *Personality and Social Psychology Bulletin* 29.5 (2003): 665-675.

59. Hancock, Jeffrey T., et al. "On lying and being lied to: A linguistic analysis of deception in computer-mediated communication." *Discourse Processes* 45.1 (2007):

1-23.

60. Bowyer, J. B. "Cheating." (1982).

61. Yuan X. Qing. “Deception detection in multi-agent systems and war-gaming”. M.S. thesis, Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire, USA, 2007.

62. Vyas, Amrish, and Lina Zhou. "On detecting deception in agent societies."*Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*. IEEE, 2005.

63. Schillo, Michael, Petra Funk, and Michael Rovatsos. "Using trust for detecting deceitful agents in artificial societies." *Applied Artificial Intelligence* 14.8 (2000): 825-848.

64. Walton, Douglas. "Deceptive arguments containing persuasive language and persuasive definitions." *Argumentation* 19.2 (2005): 159-186.

65. Miłkowski, Marcin. “What kind of cognitive process is argumentation?” Talk in Argumentation as a Cognitive Process, Toruń, UMK/PTK, May 2008.

66. Carenini, Giuseppe, and Johanna D. Moore. "Generating and evaluating evaluative arguments." *Artificial Intelligence* 170.11 (2006): 925-952.

67. Zukerman, L., Richard McConachy, and Kevin B. Korb. "Bayesian reasoning in an abductive mechanism for argument generation and analysis." *Proceedings of the national conference on Artificial Intelligence.* John Wiley & Sons LTD, 1998.

68. Shibles, Warren. "A Revision of the Definition of Lying as an Untruth Told with Intent to Deceive." *Argumentation* 2.1 (1988): 99-115.

69. Santos, Eugene, Jr., Deqing Li, and Xiuqing Yuan. "On deception detection in multi-agent systems and deception intent." *SPIE Defense and Security Symposium*. International Society for Optics and Photonics, 2008.

70. Santos, Eugene, Jr., and Deqing Li. "On deception detection in multiagent systems." *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 40.2 (2010): 224-235.

71. Li, Deqing, and Eugene Santos Jr. "Argument formation in the reasoning process: toward a generic model of deception detection." *Proceedings of the Workshop on Computational Approaches to Deception Detection*. Association for Computational Linguistics, 2012.

72. Falappa, Marcelo Alejandro, Gabriele Kern-Isberner, and Guillermo Ricardo Simari. "Belief revision and argumentation theory." *Argumentation in artificial intelligence*.

Springer US, 2009. 341-360.

73. Pasquier, Philippe, et al. "Argumentation and persuasion in the cognitive coherence theory." (2006): 223-234.

74. Pearl, Judea. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausble Inference*. Morgan Kaufmann Pub, 1988.

75. Carofiglio, V., and F. D. de Rosis. "Combining logical with emotional reasoning in natural argumentation." *3rd Workshop on Affective and Attitude User Modeling*. 2003.

76. Tenenbaum, Joshua B., Thomas L. Griffiths, and Charles Kemp. "Theory-based Bayesian models of inductive learning and reasoning." *Trends in cognitive sciences* 10.7 (2006): 309-318.

77. Mann, Thorbjoern. "The Structure and Evaluation of Planning Arguments."*Informal Logic* 30.4 (2010).

78. Buller, David B., and Judee K. Burgoon. "Interpersonal deception theory."*Communication theory* 6.3 (1996): 203-242.

79. Johnson, Marcia K., and Carol L. Raye. "Reality monitoring." *Psychological review* 88.1 (1981): 67-85.

80. Markus, Hazel. "Self-schemata and Processing Information about the Self". *Journal of Personality and Social Psychology*, 1977, 35:63–78.

81. Mehrabian, Albert. *Nonverbal communication*. Aldine De Gruyter, 1972.

82. Wiener, Morton, and Albert Mehrabian. *Language within language: Immediacy, a channel in verbal communication*. Ardent Media, 1968.

83. Tindale, Christopher William. *Acts of arguing: A rhetorical model of argument*. Vol. 101. Albany: State University of New York Press, 1999.

84. Resnick, Paul, et al. "GroupLens: an open architecture for collaborative filtering of netnews." *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM, 1994.

85. Santos, Eugene, Jr., and Gregory Johnson Jr. "Toward detecting deception in intelligent systems." *Defense and Security*. International Society for Optics and Photonics, 2004.

86. Ferreira, Ulisses. "On the foundations of computing science." *Metainformatics*. Springer Berlin Heidelberg, 2004. 46-65.

87. Li, Wentian. "Mutual information functions versus correlation functions." *Journal of statistical physics* 60.5-6 (1990): 823-837.

88. Gokhman, Stephanie, et al. "In search of a gold standard in studies of deception." *Proceedings of the Workshop on Computational Approaches to Deception Detection*. Association for Computational Linguistics, 2012.

89. Festinger, Leon. *A theory of cognitive dissonance*. Vol. 2. Stanford university press, 1962.

90. Sprites, Peter, Clark N. Glymour, and Richard Scheines. *Causation Prediction & Search 2e*. Vol. 81. MIT press, 2000.

91. Moens, Marie-Francine. *Information extraction: algorithms and prospects in a retrieval context*. Vol. 21. Springer, 2006.

92. Cowell, Robert G. "Parameter learning from incomplete data for Bayesian networks." *7th Int. workshop on Artificial Intelligence and Statistics*. 1999.

93. Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.

94. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002.

95. Beinlich, Ingo A., et al. *The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks*. Springer Berlin Heidelberg, 1989.

96. Abramson, Bruce, et al. "Hailfinder: A Bayesian system for forecasting severe weather." *International Journal of Forecasting* 12.1 (1996): 57-71.

97. Andreassen, Steen, et al. *A model-based approach to insulin adjustment*. Springer Berlin Heidelberg, 1991.

98. Andreassen, Steen, et al. "MUNIN—an expert EMG assistant." *Computer-aided electromyography and expert systems* 2 (1989): 255-277.

99. Fawcett, Tom. "An introduction to ROC analysis." *Pattern recognition letters*27.8 (2006): 861-874.

100. Crystal, David. *The Cambridge Encyclopedia of English Language*. Ernst Klett

Sprachen, 2003.

101. Weigand, Edda. "Misunderstanding: The standard case." *Journal of pragmatics*31.6 (1999): 763-785.

102. Milroy, Lesley. "Comprehension and context: Successful communication and communicative breakdown." *Applied sociolinguistics* 271 (1984): 7-31.

103. Wilhelm, Oliver. "Measuring reasoning ability." *Handbook of understanding and measuring intelligence* (2005): 373-392.

104. Asch, Solomon E. "Opinions and social pressure." (1955): 1-8.

105. Liu, Bing. "Sentiment analysis and subjectivity." *Handbook of natural language processing* 2 (2010): 568.