Group Decision and Negotiation **13**: 111–127, 2004

Criminal Identity Deception and Deception Detection in Law Enforcement

GANG WANG, HSINCHUN CHEN AND HOMA ATABAKHSH
Department of Management Information Systems, University of Arizona, 430 McClelland Hall, Tucson, AZ 85721 (E-mails: gang@bpa.arizona.edu; hchen@bpa.arizona.edu; homa@bpa.arizona.edu)

Abstract

Criminals often falsify their identities intentionally in order to deter police investigations. In this paper we focus on uncovering patterns of criminal identity deception observed through a case study performed at a local law enforcement agency. We define criminal identity deception based on an understanding of the various theories of deception. We interview a police detective expert and discuss the characteristics of criminal identity deception. A taxonomy for criminal identity deception was built to represent the different patterns that were identified in the case study. We also discuss methods currently employed by law enforcement agencies to detect deception. Police database systems contain little information that can help reveal deceptive identities. Thus, in order to identify deception, police officers rely mainly on investigation. Current methods for detecting deceptive criminal identities are neither effective nor efficient. Therefore we propose an automated solution to help solve this problem.

Key words: criminal identity deception, deception, deception detection, identity, identity fraud

1. Introduction

Criminals often intentionally falsify their identities in order to deter police investigation. Identity can be represented by a vector of key and value pairs that identify a specific person. It usually includes information such as name, gender, date-of-birth, social security number, and address. Criminal records in a local law enforcement agency show that many criminals have used deceptive identities. The identity deception problem is also encountered in the field of national security. It received increasing attention following the terrorist attacks in the United States on September 11, 2001. The FBI (2001) reported that most of the nineteen hijackers in that incident used false identities, including impersonating stolen identities and using aliases. This made it difficult for the FBI to dig out their real identities. As a result, the validity of published investigation results is still being questioned.

The 9/11 tragedy might have been avoided if investigative agents were able to recognize false identities. One reason why it is difficult for police officers to realize someone is using a stolen identity is that law enforcement agencies do not easily share information amongst themselves. One solution would be to have law enforcement agencies work together collaboratively and share information across agencies and even across countries. However, substantial efforts are required to build such a collaborative infrastructure since the existing obstacles are not only technical but also political and social. Another solution would be to address the problem of detecting criminal identity deception. For example, one

9/11 hijacker reported by the FBI (2001) used aliases such as: "Majed M.GH Moqed," "Majed Moqed," and "Majed Mashaan Moqed." Another hijacker used two dates-of-birth: "01-01-1976" and "03-03-1976." Each of these examples has some similarities that could be used to identify variations of deceptive criminal identities. For example, knowing only one date-of-birth used by a suspect, a police officer can look at variations of that date and find out whether this suspect has reported a false date-of-birth in the past.

To the best of our knowledge, criminal identity deception has not yet been addressed in any literature. In this paper we try to address this issue and propose a possible solution. In section 2 we discuss the concept of general deception and identity deception. Criminal identity deception is defined as a subset of identity deception in the law enforcement domain. We report the characteristics of criminal identity deception based on an interview with a police detective. In section 3 we describe a case study in which we investigated patterns of how criminals or suspects lie about their identities. We introduce a taxonomy of criminal identity deception built upon the case study in section 4. Methods currently employed to detect deception in law enforcement are discussed in section 5. In section 6 we propose an automated approach to detect criminal identity deception using similarity measures and approximate string matching techniques. Conclusions and future research are discussed in the last section.

2. Defining criminal identity deception

Deception has been studied in social science for many years. However, as a subcategory of deception, identity deception has not been well defined. In this section we try to define this problem based on the understanding of current deception theories. We also address the relationship between identity deception and identity theft/fraud. Criminal identity deception is the type of identity deception occurring in the law enforcement domain. To better understand this issue, we discuss its underlying causes and ways in which criminals lie about their identities.

2.1. What is deception

As a multidisciplinary concept, deception has been defined in many ways. Knapp and Comadena (1979) defined it as "the conscious alteration of information a person believes to be true in order to significantly change another's perceptions from what the deceiver thought they would be without alteration." This definition specified acts of lying as information alteration. However, the deceiver could act differently to mislead the receiver, for example, by concealing information. Mitchell (1986) defined both human and non-human deception in a general way as "a false communication that tends to benefit the communicator." One flaw in Mitchell's definition was that it implied inclusion of unconscious and mistaken deception (Vrij 2000). Ekman (1985) defined deception as "one person intends to mislead another, doing so deliberately, without prior notification of this purpose, and without having been explicitly asked to do so by the target." This definition explicitly stated

that deception is an intentional action. However, it is not necessary to exclude the prior notification of intent to deceive. Buller and Burgoon (1996) defined deception more precisely and concisely as "a sender's knowingly transmitting messages intended to foster a false belief or conclusion in the receiver." Rather than focusing on the act itself, they judged deception on the basis of the deceiver's motivations in an interpersonal communication context. This definition is also suitable for describing criminal identity deception, since criminals usually lie about their identities in an interactive environment (for example, during interrogation).

2.2. Identity and identity deception

Clarke (1994) defined human identity as "the condition of being a specified person." A specified person does not mean a specific person, but an entity described by a distinct set of characteristics. A specific person may adopt different identities at various times or maintain several at once. Identity information is a vector of distinguishing key-value pairs that include names, codes, tokens (e.g., birth certificate), knowledge (e.g., what is the person supposed to know), and biometric information (e.g., appearance, voice characteristics) (Clarke 1994).

By adopting Buller and Burgoon's definition of deception, identity deception can be defined as a sender's knowingly transmitting identity information intended to foster a false belief in the receiver. This phenomenon exists in different domains. For example, it is common to observe identity deception in a virtual community (Donath 1997), specifically the Internet.

Identity deception includes the issue of identity theft or identity fraud. As defined in the Identity Theft and Assumption Deterrence Act, identity theft refers to the action of "knowingly transferring or using, without lawful authority, a means of identification of another person with the intent to commit, or to aid or abet, any unlawful activity" (Public Law 1998). This definition focuses on the use of another person's identity and is often called impersonation. Identity deception is a broader concept than identity theft because impersonation is just one of many ways to alter an identity. For example in the hijacker case discussed previously, those hijackers deceived by using variations of their identities rather than impersonation.

2.3. Criminal identity deception in law enforcement

In this section we examine identity deception in law enforcement. Following the discussion on identity deception, we define criminal identity deception as a criminal intentionally altering his/her identity in order to foster a false belief in officers. This definition excludes the use of new identities approved by the Witness Protection Program. This program involves the legitimate acquisition of new identities through the issuance of new legitimate documentation such as birth certificates, driver's licenses, marriage licenses, homes, and so on. Subjects in this program are trying to protect themselves from being hurt rather

than to deter police investigations. Based on our definition of criminal identity deception, this is considered entirely out of the scope of this paper.

Because our topic is domain specific and requires domain knowledge, we interviewed an experienced police detective who has served in law enforcement for more than 30 years. According to him, criminals or suspects usually lie about the particulars of their identity, such as name, date of birth (DOB), address, or identification numbers, in order to deceive a police officer. Why do criminals lie? Hample (1980) concluded from an experiment that most lies were "defensive reactions to minimize trouble in situations in which lying was virtually automatic." This helps to explain the underlying causes of criminal identity deception. According to our detective expert, both suspects who committed the crime and those suspects who were not actually involved in the investigated crime will give false identities. This can be explained by Hample's statement: suspects give defensive reactions to minimize possible future troubles. However, detailed reasons need to be further studied from a psychological aspect.

In current law-enforcement computer systems, police officers run exact-match queries to locate any historical data about a suspect. When a criminal uses a false identity, even if a very similar identity is recorded in the database, an exact-match query will not bring up that record. This results in a discrepancy between available information and the need for information retrieval. According to our detective expert, criminals have found it easy and effective to escape justice by using a false identity. A felon at large may be able to escape arrest by using a falsified identity and continue to endanger society. On the other hand, it is possible that police officers may find themselves fruitlessly engaged in pursuing a most-wanted criminal who is already in a jail somewhere under a falsified identity. Both cases diminish the efficiency of police investigative activities.

Criminals usually conduct identity deception in two manners: ad-hoc and pre-planned deception. According to the detective expert, ad-hoc deception is the most common. Criminals generally attempt ad-hoc deception while they are in direct contact with police (i.e., they are being interviewed in reference to their possible involvement in a criminal activity). In certain types of crimes such as fraud or forgery, criminals may plan identity deception in advance by assuming someone else's identity or creating a false identity. In some other crimes or criminal careers (e.g., gang members), where criminals perceive that there is a strong likelihood that they will be contacted by police, they may pre-plan deception as well. Within the scope of this paper, we only address the ad-hoc manner of deception.

We have identified and defined the problem of criminal identity deception. There is still the question of how criminals lie about their identities when they commit identity deception, specifically in the ad-hoc manner. In the next section, we present findings of a case study that we conducted in a local law enforcement agency. Patterns of criminal identity deception were identified and built into a taxonomy.

3. A case study on criminal identity deception

To answer the question "in what ways do criminals lie about their identities?" we conducted a case study with criminal records from the Tucson Police Department (TPD) in order to

acquire actual patterns on our subjects. TPD has about 1.3 million criminal records kept on file. All of those records are managed by a computerized database system, which made it very convenient for us to draw samples. Records contained criminal identity information such as name, data of birth (DOB), address, identification numbers (e.g., social security numbers), race, weight, height, hair color, eye color, etc.

After examining each attribute in the records, we discarded those physical description fields that had little consequence in finding deceptive patterns. The discarded physical description fields include height, weight, hair color, and eye color. Some of these fields, such as height and weight, may change over time. According to the detective expert, those fields are considered inexact. In other fields criminals can intentionally make changes with ease. For example, hair can be dyed different colors and colored contact lenses change eye color. Therefore, these physical descriptions are too unreliable to be of any real importance. After dropping the physical description fields, we examined the remaining fields: last name, first name, address, DOB, and Social Security Number (SSN).

We then invited the police detective expert to validate the deceptive criminal identities. First, we manually located criminals having reported deceptive identities. The detective expert was asked to validate criminals' deceptive identities using his investigative methods, for example, comparing mug shots or fingerprints if available, background checking, etc. He could only confirm those deceptive identities that had strong evidence to prove that the person represented by a deceptive identity and the corresponding target criminal were the same. Strong evidence included, but was not limited to, a combination of the following: multiple matches in their previous addresses, matches in mug shots or fingerprints, matches in other types of identification number (e.g., driver's license number). We finally identified a sample of 24 criminals containing 372 criminal identities reported from incidents in which they were involved. To make clear comparisons we grouped identities by criminal. Each group had a criminal's true identity information and his/her identity variations, which were also grouped by fields like alias, DOB, SSN, and address. Gender, age, or ethnic group may or may not play a role in how people deceive. Therefore we chose an equal number of male and female criminals, ranging from 18 to 70 years in age. Also, there was an equal number of Caucasians, Hispanics, and African Americans in the selected sample (Table 1). These were the three largest ethnic groups in the City of Tucson (2001).

Table 1. The breakdown of sample records into the categories of age and ethnic groups

Ethnic group	The number of criminals in different age groups				Total
	19 and younger	20~29	30~39	40 and older	
Caucasian	2	2	2	2	8
Hispanics	2	2	2	2	8
African American	2	2	2	2	8
Sum:					24

4. A taxonomy of criminal identity deception

Patterns of criminal identity deception were noticed when we compared an individual's deceptive records to his/her real identity record. We categorized criminal identity deception into four types: name deception, residency deception, DOB deception, and ID deception. A taxonomy of criminal identity deception was built based upon our observations and analyses.

4.1. Name deception

As noted by our police detective expert, it is common for criminals to lie about their names. In our case study, all twenty-four criminals identified had used a false name. The name field contains three parts: last name, first name, and middle initial. Our sample showed that criminals were more likely to make variations on name spelling and/or name sequence than to use a completely different name. Each type of name deception is described below.

4.1.1. Spelling variation

Deception by spelling variation was the most frequent type in the category of name deception. It can be further categorized based on the way criminals make spelling changes. A criminal could use several types of spelling variations simultaneously.

(1) Name with similar pronunciation:

This type of spelling variation means that either the first name or the last name is replaced by a name phonetically the same or similar but different in spelling. To illustrate, one subject named "Cecirio" used the false name "Cicero" instead. In our sample, this type of name deception was found in ten out of twenty-four criminals (41.7%).

(2) Abbreviations and add-ons:

In this type of spelling deception, criminals may use abbreviated names or add additional letters to their real ones. A good example of an abbreviation is using "Ed" instead of "Edward." Similarly, using "Edwardo" instead of "Edward" is a good example of Add-on. Seven criminals in our sample data (29.2%) were found to have used this type of deception.

(3) Changing middle initial:

Unlike first and last names, spelling is not an issue for middle initials. Criminals sometimes just simply change them. In our sample data, ten criminals (62.5%) either left out or changed their middle initials, or fabricated a middle initial when there was none. The middle initial is not as important as a first name or a last name because when a police detective is conducting a name search on a specific suspect in the database the middle initial is always ignored. According to the detective expert, police officers only use the middle initial to differentiate between suspects when there are several records for a common name.

4.1.2. Completeness and sequence

Our sample showed two types of deception under this category: name swap and partly missing name.

(1) Name swap:

Name swap is defined as the action of transposing one's first and last names. For example, "Edward Alexander" can be altered as "Alexander Edward." This type of deception can only happen where transposing the first and last names does not raise immediate doubt. People consider some names to be used as first names or last names only. For instance, "Smith" is usually a last name and could be suspected immediately if a criminal reports it as his first name. In our sample of criminal records, two criminals (8.3%) were found to have used name swap.

(2) Partly missing name:

Partly missing name is defined as the situation in which a criminal record lacks either the first name or the last name. Seven criminals in our sample data (29.2%) had used this type of deception. Criminals might not report part of their names intentionally in order to interfere with the investigation against them. On the other hand, it is also possible that police officers lose part of the information during the data entry or they are unable to acquire complete information during the investigation.

4.1.3. Completely deceptive name

A completely deceptive name means that a criminal uses a name, either a first name or a last name or a full name, which is totally different from and irrelevant to its real representation. In that case we would not be able to see any of the patterns described previously. Seven criminals (29.2%) were deceptive in this way. For example, a subject named "Joy Baker" falsified her first name as "Rebecca Baker." According to our police detective expert, criminals in this case usually choose the name of a brother, sister, or partner.

4.2. Residency deception

Residency deception is related to address information. Generally, an address is composed of a street number, a street direction, a street name, and a street type, e.g., "1201 W Highland Ave." Suspects usually make changes to only one portion of the full address. Eight criminals in our study were found to have used a deceptive address. Based on our observation, street number was the most commonly altered part.

4.2.1. Deception on street number

Among the eight criminals we examined, seven (87.5%) deceived on the street number. Deception was made by changing, removing, or inserting some digits into the real street number. For example, the address "1201 W. Highland Ave." can be altered as "1211 W. Highland Ave.," "120 W. Highland Ave.," or "11201 W. Highland Ave." In most cases, there were no more than two digits altered.

4.2.2. Deception on street direction

A street direction can be longitudinal (e.g., North, South) or latitudinal (e.g., West, East). Most streets have only one type of direction. Some streets may have both types when swerves exist. For streets having only one type of direction, criminals who alter the direction are very likely to change other portions of the address such as street type and street name as well. That will make the deception more reasonable. However, one criminal in our study only altered the direction without making other changes, which simply created a non-existing address. Another criminal altered both street direction and street type, which made it a valid address and hard to detect. For streets having both types of direction, criminals may simply alter the direction to make the false address valid. One subject in our case study changed the street direction from "East" to "South" with other portions of his address intact, which exemplified that kind of deception.

Three criminals (37.5%) were found to have altered street directions. Two criminals created a valid false address while the other created an invalid one.

4.2.3. Deception on street name

A street name can be numeric (e.g., 2nd St.) or textual (e.g., Hatfield St.). In our case study, three subjects (37.5%) falsified their street names. One criminal altered his numeric street name to another numeric name. He reported "73 E. 34th St" as "73 E. 35th St." The other two criminals used deception on their textual street names by making spelling variations. One reported "Calle Arroyito" as the street name instead of "Calle Del Arroyito," while the other reported "Desert" instead of "Desert Mesa." This type of deception is similar to name deception and most types of variation discussed in name deception are expected to occur in the deception on street names as well.

4.2.4. Deception on street type

Values for street type can be entities such as Street, Road, Avenue, Drive, Boulevard, and Way. Two criminals (25%) were found to deceive on street types. In both cases criminals also altered other parts of the address. It seems that criminals alter street types in order to make the false address look valid. For example, the real address for a criminal was "144 E. 9th St." In one incident report, he altered his address to "144 S. 9th Ave." If he had not altered the street type, the address would have been an invalid one because there is no address such as "S. 9th St." in the Tucson area. This is still an ad-hoc manner of deception.

4.3. DOB deception

DOB deception is the most common type of criminal identity deception. Our case study showed that sixteen criminals (66.7%) in the sample had deceived on DOB. The DOB field in the TPD database has an eight-digit number representing the year, the month, and the day respectively. For example, "19700215" represents a DOB of February 15th, 1970. By studying the deceptive cases, we found that suspects usually made only slight changes to

their deceptive DOB. For example, "19700215" might have been falsified as "19700205" by changing the day. Changes can also be made to month and year. In all DOB deception cases in our sample, 65% only falsified one portion of their DOB, 25% made changes on two portions of their DOB, and 10% made changes to all three portions.

The way criminals altered their DOB was similar to name deception. Criminals made "spelling variations" to their real DOB, such as replacing a couple of digits with false digits and transposing digits. Thus, if two records in the database show the same or similar names and two different but very similar DOB's (for example, 19560608 and 19560806), a police officer can deduce that both records are in fact the same person.

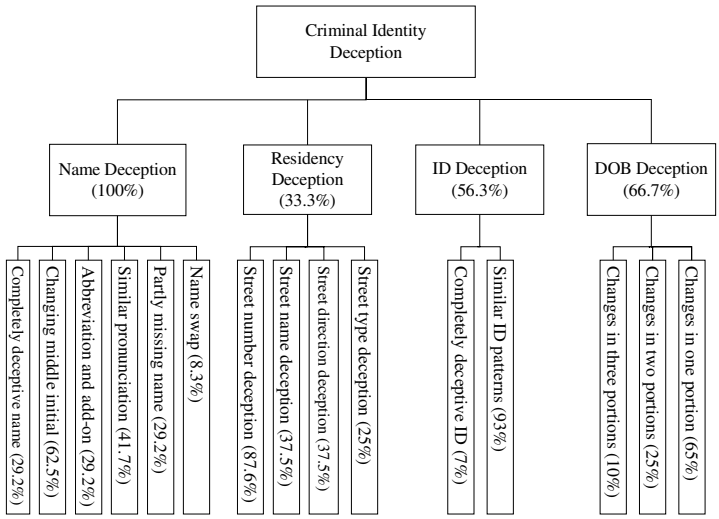
4.4. ID deception

ID is a unique sequence of numbers and letters that is associated with an identification document In the US an identification document can be a passport, a driver's license, birth certificate, or a social security card. The law enforcement database system we studied stores Social Security Numbers (SSN) along with other types of ID (e.g., driver's license number). There were many more records having SSNs than other types of ID. Therefore we chose the SSN to represent all ID patterns. We assume the deception patterns occurring in SSNs are the same as those occurring in any other types of ID.

SSN is a nine-digit number, which is a unique identification number for each person. One may not have a SSN under some circumstances (e.g., a non-citizen staying for nonwork purposes). In this study, we only used those records having SSNs. Within our sample, 56.3% of the suspects used a falsified SSN. Among those cases of SSN deception, most of them (96%) varied no more than two digits from the corresponding correct ones. One example of SSN deception is the ID "123-45-6789" may be changed to "123-46-6789" or "123-35-9789." Still, criminals may make variations similar to those described for name deception, such as number swap and completely deceptive numbers. In our sample, we found one case where a criminal gave a totally different SSN.

Figure 1 summarizes the different types of criminal identity deception described above. During the case study we noticed it was sometimes difficult to distinguish between deceptive records and records having data entry errors. In this case, we examined all fields in the suspected identity. When there were more altered fields the record was more likely to be deceptive. For example, partly missing names may result from intentional criminal deception or from operational information loss. We identified this type of name deception by comparing records in the database and checking for the completeness of names. If two different records showed similar names with one being more complete than the other (e.g., the first name missing in one record), then we looked at other fields in those two records (e.g., DOB, SSN, physical characteristics). If other fields were very similar we deduced that those two people were in fact the same person and we considered this to be a case of deception. In the same manner if two names were similar with the exception of the first and last names being swapped, we compared other fields in those two records in order to see whether we could deduce that this was a deceptive case.

(a)



(b)

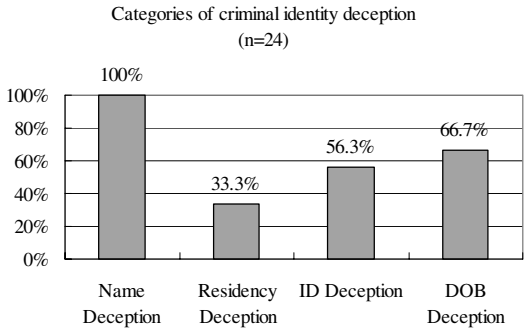
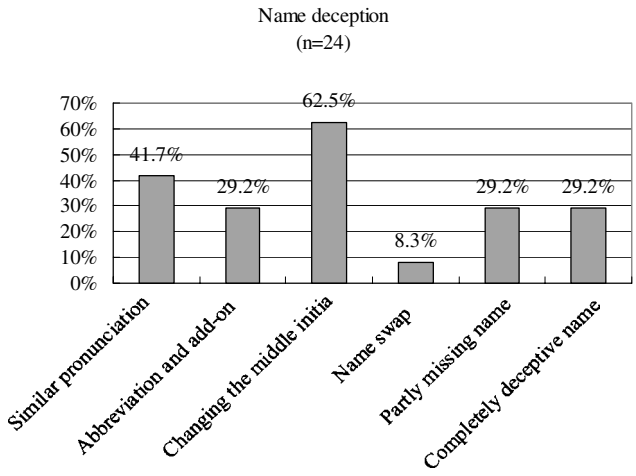


Figure 1. (a) A taxonomy of criminal identity deception. (b) Bar-chart of the major categories of criminal identity deception. (c) Bar-chart of the types of name deception. (d) Bar-chart of the types of residency deception. Each criminal may have more than one type of name deception.

5. Methods to detect criminal identity deception

Police detectives usually do not specifically aim to detect criminal identity deception. A false criminal identity is often revealed as a byproduct of other investigation activities, unless there are serious doubts about a criminal's identity. There are techniques, such as observing a combination of physical, emotional and mental symptoms of deception, developed for deception detection in law enforcement (Aubry and Caputo 1980). However, those techniques are typically used to verify statements made by criminals. None of them is specifically designed for revealing lies about identities.

(c)



(c)

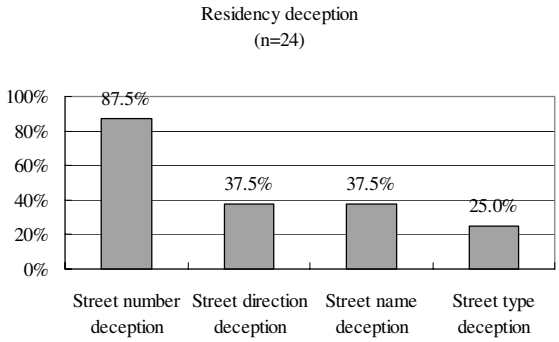


Figure 1. Continued.

Vrij (2000) summarized three ways for detecting lies in law enforcement. The first method is to observe liars' non-verbal behavior, such as their body movements (e.g., scratching the head), their emotional expressions, their facial expression (e.g., blinking of the eyes), and vocal characteristics (e.g., pitch of voice). It has been shown that there are automatic links between emotions and non-verbal behaviors (Ekman 1992). Non-verbal cues to deception are more likely to occur if the lie is difficult to fabricate (Vrij 2000). The emotional fluctuation caused by the action of lying will influence one's behavior, which could expose deception. The second method is to analyze verbal characteristics of what a subject said. Vrij (2000) defined several types of verbal characteristics including negative statements, plausible answers, irrelevant information, over-generalized statement, self-references, direct answers, and response length. Verbal cues can help to discriminate between

deceptive and truthful statements in the sense that some verbal criteria are more likely to occur in false rather than in truthful statements. Statement Validity Assessment (SVA) and Reality Monitoring are two popular techniques for detecting verbal cues. The third way is to examine physiological responses such as blood pressure, heart rate, palmar sweating, respiration, and so on. The device that can detect physiological activities is called a polygraph. These three techniques are used by people specifically trained for that purpose (for example, polygraph examiners) (Vrij 2000). The practical accuracy of those three methods, as used by experts in detecting lies, is reported in Table 2.

In practice, police officers and detectives generally perform worse than the trained experts (DePaulo and Pfeifei 1986; Ekman and O'Sullivan 1991; Kohnken 1987; Kraut and Poe 1980), so that many cases of deception are not discovered. Deception can also be revealed by investigation (e.g., checking a criminal's history information), which is time-consuming and involves great amounts of manual information processing. For example, a crime analyst may find a suspect using a false identity when he/she is investigating crime patterns and trends for a particular case by conducting link analysis. This method is often used to construct criminal networks from database records or textual documents. Sometimes police detectives compare the suspect's identity information to the current criminal records in the database to find discrepancies that indicate deception, which is simple but also time-consuming. It is also unrealistic for a human detective to examine all records in the database one by one, because of the huge amount of criminal records in a law enforcement database system (for example, the TPD has about 1.3 million records currently).

6. Proposed approach

6.1. An algorithmic approach

The experimental results listed in Table 2 show that the three methods introduced by Vrij are not reliable enough to detect lies. Although widely used, polygraph results are not even admissible as evidence in the Supreme Court of the United States (CNN 1997). These meth-

Table 2. Accuracy report of the three methods for detecting lies (Vrij 2000)

	Non-verbal	Verbal	Physiological
How good are experts at detecting lies?	51–82% accuracy in truth detection 30–66% accuracy in lie detection 31–64% total accuracy	76% accuracy in truth detection* 68% accuracy in lie detection*	72% accuracy in truth detection** 87% accuracy in lie detection** 96% accuracy in truth detection 59% accuracy in lie detection

*Results on average with the Criteria-Based Content Analysis (CBCA).
**Results on average with the Control Question Technique.
***Results on average with the Guilty Knowledge Test.

ods can indicate lies to some extent, but they fail to extract the true information linked to the lies. According to our police expert, those general approaches can hardly apply to the previously defined problem of criminal identity deception. Identity information contains no verbal cues that can be used in verbal characteristic analysis. Usually, a criminal or a suspect's identity is reported at the time they are interviewed in the field by a police officer. Non-verbal cues usually are too subtle to be noticed by the officer.

By examining the characteristics of criminal identity deception, we found clues that might help provide a solution to this problem. First, values in criminal identity information are textual strings, which consist of both letter and numerical characters. For example, names are comprised of letters while birth dates are represented as 8 numerical digits. Second, criminals deceive by altering their identities. In most cases observed in the study, criminals "wisely" changed only a small portion of their original identity information. The alteration showed similar patterns to the corresponding true identity. For example, spelling variations and approximate pronunciation are two commonly used tricks to alter names. Third, different fields in identity information need to work together in order to reveal an identity deception. In most circumstances, finding one deceptive field cannot adequately indicate a deception case. For example, there is one record for "Tom Smith" and another one for "Tommy Smith," and both records are completely different in other fields such as address, birth date, and identification number. "Tommy Smith" could be a deceptive name for "Tom Smith," but it is most likely that the two records represent two different people when the other identity fields are considered. On the other hand, if all other fields exactly match, a difference in name is probably caused by data entry error.

Approximate string matching techniques have been studied in computer science for years (Navarro 2001) and may detect spelling differences between strings or search for a given string in a text allowing spelling variations. Based on what we have discovered in the case study, such a technique may help us to locate falsified information (e.g., name deception) if given the original. For example, edit distance is one of the well-known approximate string matching algorithms (Levenshtein 1966). It calculates the smallest number of character insertions, deletions, and substitutions required to change one string into another. As we built the taxonomy of criminal identity deception, we observed that most of the altered identity information involved character insertions, deletions, and substitutions (e.g., name add-ons and abbreviations). We expect that the edit distance between the original information and the altered one can indicate their association to some extent.

Based on the above analysis, we propose a similarity-based approach to automatically find a link between true identity information and corresponding deceptive identity information. We can do so by first assuming that all criminal identity records currently existing in police databases are true. By comparing an identity input with each record in police databases using our proposed approach, we will consider it a deception if there is a similar but different record in databases. We have to be careful in defining the "similar but different" relationship between two identity records. "Different" means that two identities are not exactly the same, while "similar" indicates that two identities actually represent the same person. However, we cannot assume all current records are true simply because they exist in the police databases. It is possible that a criminal has given deceptive identities beginning with his/her first record. In this case, we cannot simply tell whether a new input is

deceptive or not by finding its corresponding similar and different records in police databases. There will be three possibilities for this case: the new input is true and the existing record is deceptive, the existing record is true and the new input is deceptive, both identities are deceptive. In practice, according to our detective expert, it is more important to find relevant information than to ascertain the truthfulness of an identity. Therefore we should broaden the meaning of deception detection for criminal identity deception in law enforcement. Rather than determining whether a specific identity is deceptive or not, we aim to find at least one deceptive identity in an identity pair and provide more relevant information to assist police investigation.

We propose to use string comparison techniques in measuring the similarity between pieces of textual information. The fields observed in the case study all contain textual values, consisting of letters and numerical digits. String comparison techniques, such as edit distance (Levenshtein 1966) and Soundex, can detect spelling and pronunciation differences between two strings and give a numerical measure representing the similarity or dissimilarity between them. The overall similarity measure between two identity records is concluded based on the comparisons of individual fields:

$$S = S_n + S_a + S_d + S_i$$

where S, S_n, S_a, S_d, S_i represent the overall similarity measure, name similarity, address similarity, DOB similarity, and ID similarity respectively.

6.2. Experimental results

We tested the feasibility of the proposed algorithm with a set of real criminal records drawn from the Tucson Police Department (TPD) (Wang et al. 2003). With the help of the same police detective expert, we chose 120 criminal records with identified deceptions. Each record had complete information in the four fields used by our proposed algorithm: name, address, DOB, and SSN. This sample set involved 44 criminals each of whom had an average of 3 records. The sample record set was used to train and test our proposed algorithm so that records pointing to the same suspect could be associated with each other. It was validated by a standard hold-out sampling method: 80 records (2/3) were used for training, while the remaining 40 records (1/3) were used for testing purposes.

During the training process, we normalized them using a Euclidean distance function in order to make similarity measures consistent. We tried different threshold values to classify similarity measures between records into matched (deceptive) and unmatched pairs. When the threshold value was set to 0.48, the algorithm achieved its highest accuracy of 97.4% in linking deceptive records pointing to the same person. We applied this optimal threshold value to the record set for testing. The result showed that the accuracy of linkage in the testing data was 94%. This experiment showed that our proposed algorithm is feasible and effective in associating deceptive records pointing to the same suspect.

7. Conclusion and discussion

In this paper we have defined criminal identity deception based on an understanding of the various theories of deception. Based on an interview with a police detective, we discussed the aspects of criminal identity deception in a practical context. In a case study conducted in a local law enforcement department (TPD) we found different types of criminal identity deception. A taxonomy was built based on deception patterns revealed through the case study. We explored some generic methods that are currently employed to detect deception in law enforcement. However, these methods are neither effective nor efficient in detecting criminal identity deceptions. Based upon the deception patterns identified in the taxonomy, we proposed a similarity-based approach to automatically find a link between true identity information and corresponding deceptive identity information. A preliminary experiment showed that the proposed algorithm was effective in associating deceptive records pointing to the same suspect.

7.1. Implications for automating deception detection

Law enforcement agencies require an effective and efficient method for detecting criminal identity deceptions. Exact queries based on deceptive information and general deception detection techniques are not effective for solving this problem. Police officers would effectively uncover deceptive identities if they could compare a suspect's identity to each criminal history record in police databases. However, the huge number of criminal history records prevents them from doing so. Having studied the various types of criminal identity deception, we propose an algorithmic approach that can inherit the heuristics for detecting deceptive identities by comparisons, from an officer. This approach, expected to be as effective as the manual approach, has to be automated in order to efficiently cope with millions of criminal history records. Moreover, an automated deception detection approach can perform more effectively than experienced police officers because as the number of comparisons increases, a human expert can easily get physically exhausted and make mistakes while an algorithmic approach can perform well constantly. In this paper, we have shown the effectiveness (detection accuracy) of our proposed algorithm with preliminary experimental results using a small dataset. Efficiency evaluation is one of our future tasks as well as the effectiveness evaluation when using a large dataset.

7.2. Limitations

The proposed algorithm design has several limitations. First, it assumes that each attribute in a criminal record has a value, and does not take into account data quality problems such as missing attribute values. However, generally there are few complete records (i.e., records having no missing values) in databases. For example, only 24% of criminal records in the TPD have values in all the four fields used by our proposed algorithm. Second, such an algorithmic approach can only detect the types of criminal identity deceptions defined in

our taxonomy. It will not work with identity theft/fraud problems, in which a criminal intentionally uses another person's identity because it is highly unlikely that the criminal's real identity would be similar to the impersonated one. Third, it is possible that the algorithm considers discrepant records resulting from data entry errors rather than the criminal's intentional act of deception. We can only argue that if there are more than one altered fields, the record is more likely to be deceptive than to have been a data entry error. However, we cannot rule out the possibility that all fields in the record have been altered by data entry errors, although the likelihood of this is small.

7.3. Future work

The proposed automated deception detection algorithm needs further evaluation in terms of efficiency and effectiveness for large datasets. Considering the large proportion of missing values contained in real police databases, we aim to extend the algorithm to handling criminal identity records with missing values.

An automated deception detection system based on the proposed algorithm will be fully developed and incorporated into our ongoing COPLINK project (Hauck et al. 2002), which has been under development at the University of Arizona's Artificial Intelligence Lab, in collaboration with the Tucson Police Department (TPD) and the Phoenix Police Department (PPD), since 1997.

Acknowledgements

This project has primarily been funded by the following grants: NSF, Digital Government Program, "COPLINK Center: Information and Knowledge Management for Law Enforcement," #9983304, July, 2000-June, 2003; National Institute of Justice, "COPLINK: Database Integration and Access for a Law Enforcement Intranet," July 1997–January 2000. We would also like to thank the Digital Equipment Corporation External Technology Grants Program, agreement #US-1998004, for its award of an equipment grant of a DEC Alpha Server for the COPLINK Project. We would like to thank the following people for their support and assistance: Lt. Jennifer Schroeder, Detective Tim Petersen, Dan Casey and other personnel from the Tucson Police Department; Members of the University of Arizona Artificial Intelligence Lab and especially the COPLINK team; and Sarah Marshall, English editor.

References

Aubry, A. S. Jr. and R. R. Caputo. (1980). *Criminal Interrogation*, 3rd Ed., Charles C Thomas Publisher.

Buller, D. B. and J. K. Burgoon. (1996). "Interpersonal Deception Theory." *Communication Theory* 6, 203–242.

Buller, D. B., and J. K. Burgoon. (1998). "Emotional Expression in the Deception Process," *in* P. A. Andersen and L. K. Guerrero (eds.), *Handbook of Communication and Emotion*. San Diego, CA: Academic Press, 381–402.

Burgoon, J. K., D. B. Buller, L. K. Guerrero, W. Afifi, and C. Feldman. (1996). "Interpersonal Deception: XII. Information Management Dimensions Underlying Deceptive and Truthful Messages," *Communication Monographs* 63, 50–69.

City of Tucson, Department of Transportation (2001). "Ethnic Groups in the City of Tucson," http://dot.ci.tucson.az.us/ttdot/dsisurvey/sld011.htm.

Clarke, R. (1994). "Human Identification in Information Systems: Management Challenges and Public Policy Issues," *Information Technology & People* 7 (4), 6–37.

CNN, "Supreme Court to Hear Arguments on Lie Detectors," http://www.cnn.com/US/9711/03/scotus.polygraph/index.html.

DePaulo, B. M. and R. L. Pfeifei. (1986). "On-the-job Experience and Skill at Detecting Deception," *Journal of Applied Social Psychology* 16, 249–267.

Donath, J. S. (1988). "Identity and Deception in the Virtual Community," *Communities in Cyberspace*. Routledge.

Ekman, P. (1985). *Telling Lies: Clues to Deceit in The Marketplace, Politics and Marriage*. New York: W. W. Norton.

Ekman, P. (2001). *Telling Lies: Clues to Deceit in The Marketplace, Politics and Marriage*, 3rd Ed., New York: W. W. Norton.

Ekman, P. and M. O'Sullivan. (1991). "Who Can Catch a Liar?" *American Psychologist* 46 (9), 913–920.

Federal Bureau of Investigation (2001). "September 11 Hijackers: Names and Photographs on FBI.GOV." U.S. Department of Justice, Washington D. C., http://www.fbi.gov/pressrel/pressrel01/092701hjpic.htm.

Hample, D. (1980). "Purposes and Effects of Lying," *Southern Speech Communication Journal* 46, 33–47.

Hauck, R.V., H. Atabakhsh, P. Ongvasith, H. Gupta, and H. Chen. (2002). "Using COPLINK to Analyze Criminal-Justice Data," *IEEE Computer*, March 2002.

Knapp, M. L. and M. E. Comadena. (1979). "Telling It Like It Isn't: A Review of Theory and Research on Deceptive Communication," *Human Communication Research* 5, 270–285.

Kohnken, G. (1987). "Training Police Officers to Detect Deceptive Eyewitness Statements: Does it work?" *Social Behavior* 2, 1–17.

Krauss, R. M. (1981). "Impression Formation, Impression Management, and Nonverbal Behaviors," *in* E. T. Higgins, C. P. Herman, and M. P. Zanna (eds.), *Social Cognition: The Ontario Symposium*, Vol. 1. Hillsdale, NJ: Erlbaum, 323–341.

Kraut, R. E. and D. Poe. (1980). "On the Line: The Deception Judgments of Customs Inspectors and Laymen," *Journal of Personality and Social Psychology* 39, 784–798.

Levenshtein, V. L. (1966). "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," *Soviet Physics Doklady* 10, 707–710.

Mitchell, R. W. (1966). "A Framework for Discussing Deception," *in* R. W. Mitchell and N. S. Mogdil (eds.), *Deception: Perspectives on Human and Nonhuman Deceit*. Albany: State University of New York Press, 3–4.

Public Law. "Identity Theft and Assumption Deterrence Act." As amended by Public Law 105–318, 112 Stat. 3007 (Oct. 30, 1998) http://www.ftc.gov/os/statutes/itada/itadact.htm#003.

Navarro, G. (2001). "A Guided Tour to Approximate String Matching," *ACM Computing Surveys* 33 (1), 31–88.

Vrij, A. (2000). *Detecting Lies and Deceit: The Psychology of Lying and The Implication for Professional Practice*. John Wiley & Sons, Ltd.

Vrij, A. and F. W. Winkel. (1993). "Objective and Subjective Indicators of Deception," *in* N. K. Clark and G. M. Stephenson (eds.), *Children, Evidence, and Procedure*. Leicester, UK: The British Psychological Society.

Wang, G., H. Chen, and H. Atabakhsh (2003). "Automatically Detecting Deceptive Criminal Identities," *Communications of the ACM* (to appear).