by Gang Wang, Hsinchun Chen,
and Homa Atabakhsh

AUTOMATICALLY DETECTING DECEPTIVE CRIMINAL IDENTITIES

Fear about identity verification reached new heights since the terrorist attacks on Sept. 11, 2001, with national security issues related to detecting identity deception attracting more interest than ever before. Identity deception is an intentional falsification of identity in order to deter investigations. Conventional investigation methods run into difficulty when dealing with criminals who use deceptive or fraudulent identities, as the FBI discovered when trying to determine the true identities of 19 hijackers involved in the attacks. Besides its use in post-event investigation, the ability to validate identity can also be used as a tool to prevent future tragedies.

Here, we focus on uncovering patterns of criminal identity deception based on actual criminal records and suggest an algorithmic approach to revealing deceptive identities.

Interpersonal deception is defined as a sender knowingly transmitting messages intended to foster a false belief or conclusion by the receiver [1]. Methods have been developed to detect deception using physiological measures (for example, polygraph), nonverbal cues, and verbal cues. Nonverbal cues are indications conveyed through communication channels such as micro-expression (for example, facial expression), eye movement, and body language. Verbal cues are linguistic patterns exhibited in messages that may include deception. The veracity of verbal cues can be measured

THE CRIMINAL MIND IS NO MATCH FOR SOME OF THE LATEST TECHNOLOGY DESIGNED TO DETERMINE FACT FROM FICTION IN SUSPECT IDENTITIES.

ILLUSTRATION BY TERRY MIURA

WE FOCUS ON UNCOVERING PATTERNS OF CRIMINAL IDENTITY DECEPTION BASED ON ACTUAL CRIMINAL RECORDS AND SUGGEST AN ALGORITHMIC APPROACH TO REVEALING DECEPTIVE IDENTITIES.

by empirical techniques (for example, Statement Validity Assessment and Criteria-Based Content Analysis) [7]. Police officers are trained to detect lies by observing nonverbal behaviors, analyzing verbal cues, and/or examining physiological variations. Some are also trained as polygraph examiners. Because of the complexity of deception, there is no universal method to detect all types of deception. Some methods, such as physiological monitoring and behavioral cues examination, can only be conducted while the deception is occurring. Also, there is little research on detecting deception in data where few linguistic patterns exist (for example, profiles containing only names, addresses, and so on). Therefore, existing deception detection techniques developed for applications in communication and physiology are not suitable for discovering deception in identity profiles.

It is a common practice for criminals to lie about the particulars of their identity, such as name, date of birth, address, and Social Security number, in order to deceive a police investigator. For a criminal using a falsified identity, even if it is one quite similar to the real identity recorded in a law enforcement computer system, an exact-match query can do very little to bring up that record. In fact, criminals find it is easy and effective to escape justice by using a false identity.

A criminal might either give a deceptive identity or falsely use an innocent person's identity. There are currently two ways law enforcement officers can determine false identities. First, police officers can sometimes detect a deceptive identity during interrogation and investigation by repeated and detailed questioning, such as asking a suspect the same question ("What is your Social Security number?") over and over again. The suspect might forget his or her false answer and eventually reply differently. Detailed questioning may be effective in detecting lies, such as when a suspect forgets detailed information about the person whose identity he or she is impersonating. However, lies are difficult to detect if the suspect is a good liar. Consequently, there are still many deceptive records existing in law enforcement data. Sometimes a police officer must interrogate an innocent person whose identity was stolen, until the person's innocence is proven.

Second, crime analysts can detect some deceptive identities through crime analysis techniques, of which link analysis is often used to construct criminal networks from database records or textual documents. Besides focusing on criminal identity information, link analysis also examines associations among criminals, organizations, and vehicles, among others. However, in real life crime analysis usually is a time-consuming investigative activity involving great amounts of manual information processing.

Record Linkage Algorithm

A literature survey was conducted to identify research that could contribute to our understanding of criminal profile analysis. In his review of this field, Winkler [8] defined record linkage as a methodology for bringing together corresponding records from two or more files or for finding duplicates within a file. Record linkage originated from statistics and survey research. Newcombe [5] pioneered this work in a study designed to associate a birth record in a birth profile system with a marriage record in a marriage profile system if information in both records pointed to the same couple. His work enabled the first computerized approach to record linkage. In recent years, record linkage techniques have incorporated sophisticated theories from computer science, statistics, and operations research [8]. Work on library holdings duplication is also a related field.

Two basic components in record linkage are the *string comparator* and the *weight determination* method. The string comparator can determine the degree of agreement between corresponding attributes, such as names, in two records. The weight determination is a mechanism to combine agreement values of all fields and of results in an overall degree of agreement between two records. The performance of a string comparator is very important because it is the key component in computing agreement values. Although current string comparator methods employed in record linkage have different limitations, they can be improved significantly for various applications.

Phonetic string comparator. To compute agreement values between surnames, Newcombe [5] encoded surnames using the Russell Soundex Code, which represented the phonetic pattern in each surname. According to the rules of Soundex coding, surnames were encoded into a uniform format having a prefix letter followed by a three-digit number. Surnames having the same pronunciation in spite of spelling variations should produce identical Soundex codes. For example, “PEARSE” and “PIERCE” are both coded as “P620.” However, Soundex does not work perfectly. In some cases, names that sound alike may not always have the same Soundex code. For example, “CATHY” (C300) and “KATHY” (K300) are pronounced identically. Also, names that do not sound alike might have the same Soundex code; for example, “PIERCE” (P620) and “PRICE” (P620).

A pair of strings		Soundex	Jaro’s	Edit distance
JONES	JOHNSONS	0.75	0.79	0.50
MASSEY	MASSIE	1.00	0.889	0.66
SEAN	SUSAN	0.50	0.783	0.60
HARDIN	MARTINEZ	0.75	0.00	0.50
JON	JAN	1.00	0.00	0.66

Table 1. Comparison between Soundex, Jaro’s method, and Edit distance.

Spelling string comparator. A spelling string comparator compares spelling variations between two strings instead of phonetic codes. In another pioneering record linkage study, Jaro [3] presented a string comparator dealing with typographical errors such as character insertions, deletions, and transpositions. This method has a restriction in that common characters in both strings must be within half of the length of the shorter string.

String comparison, whether string distance measures or string matching, has also attracted the interest of computer scientists. A common measure of similarity between two strings is defined by Levenshtein as “edit distance” [4], that is, the minimum number of single character insertions, deletions, and substitutions required to transform one string into the other. The edit distance measure outperforms Jaro’s method because it can deal with all kinds of string patterns. Since edit distance is designed to detect spelling differences between two strings, it does not detect phonetic errors.

Porter and Winkler [6] showed the effect of Jaro’s method and its several enhanced methods on last names, first names, and street names. In order to compare the Soundex coding method, Jaro’s method, and edit distance, we calculated several string examples (used in [6]) using Soundex and edit distance respectively. Table 1 summarizes a comparison of the results from Soundex, Jaro’s method, and edit distance. Each number shown in the table represents a similarity measure (a scale between 0 and 1) between the corresponding strings. We noticed that Soundex measures gave improper ratings when two strings happened to be encoded similarly, such as “JONES” (J520) and “JOHNSONS” (J525), “HARDIN” (H635) and “MARTINEZ” (M635). Edit distance measures were capable of reflecting the spelling differences in cases where Soundex measures were improper. Jaro’s method could also detect spelling variations between strings. However, it was unable to compare certain string patterns (with scores of zero). In order to capture both phonetic and spelling similarity of strings, a combination of edit distance and Soundex was selected for our research.

A Taxonomy of Criminal Identity Deception

In order to identify actual criminal deception patterns, we conducted a case study on the 1.3 million records at Tucson Police Department (TPD). Guided by a veteran police detective with over 30 years of service in law enforcement, we identified and extracted 372 criminal records involving 24 criminals—each having one real identity record and several deceptive records. The 24 criminals included an equal number of males and females, ranging in age from 18 to 70. Records contained criminal identity information, such as name, date of birth (DOB), address, identification numbers, race, weight, and height. Various patterns of criminal identity deception became apparent when we compared an individual’s deceptive records to his or her real identity record.

Or discarded physical description attributes (for example, height, weight, hair color, eye color) that had little consequence for deception detection. With visual scrutiny, suspects apparently do not lie about their height or weight. Eye color and hair color are too unreliable to be of any real importance. Criminals can

We present a record-linkage method based on string comparators to associate different deceptive criminal identity records. The experimental results have shown the method to be promising.

easily make changes to those attribute.

In the remaining attributes we found different patterns of deception in each one. Consequently, we categorized criminal identity deception into four types: name deception, residency deception, DOB deception, and ID deception. The taxonomy of criminal identity deception was built upon the case study and is summarized in Figure 1.

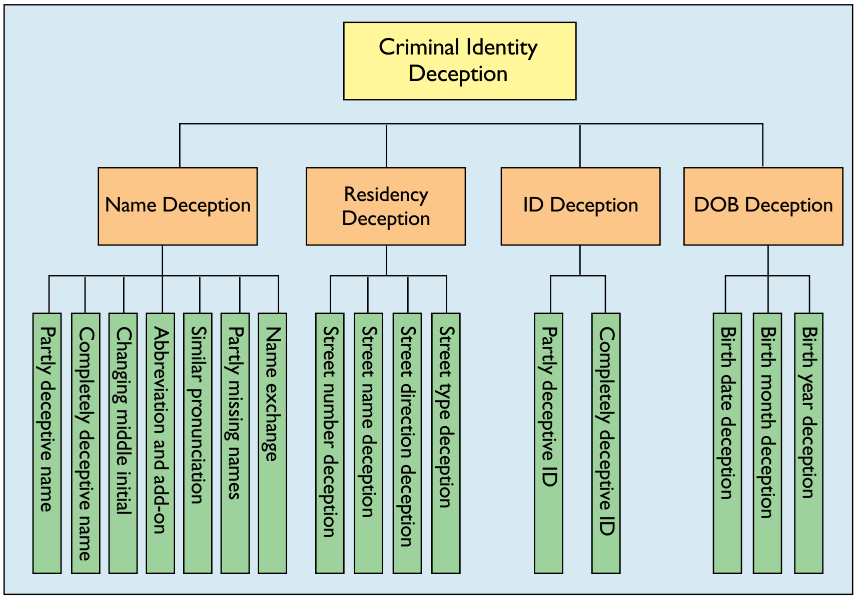

Figure 1. Taxonomy of criminal identity deception.

Name deception can take on a variety of options:

- *Partly deceptive name:* 62.5% of the criminal records in the sample data set had more than once given either a false first name or a false last name. For example, "Ed Garcia" might have been changed to "Ted Garcia."
- *Using a completely different name:* 29.2% of the records had completely false names. Both first name and last name were false.
- *Changing middle initial:* Instead of a full middle name, only middle initials are shown in the police profiles. 62.5% of the records had modified middle initials, while the first name and last name remained intact. Criminals either left out or changed their middle initials. Also, they sometimes fabricated a middle initial when there was none.
- *Abbreviation and add-on:* 29.2% of the criminal records had abbreviated names or additional letters added to their real names. An example of this is using "Ed" instead of "Edward," or "Edwardo" instead of "Edward."
- *Similar pronunciation:* This means using a deceptive name having the same or similar pronunciation, but spelled differently. In our sample, 42% of the criminals used this method of deception. For example, "Cecirio" can be altered to "Cicero."
- *Name exchange:* 8% of the criminals transposed last and first names. For example, "Edward Alexander" might have become "Alexander Edward."

DOB deception is easier than name deception to define simply because it consists of year, month, and day. By studying the deceptive cases, we found that suspects usually made only slight changes to their DOBs. For example, "02/07/70" might have been falsified as "02/08/70." Changes to month or year also were frequent in the sample. 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.

ID deception. A police department uses several types of identification numbers, such as Social Security number (SSN) or FBI ID number if one is on record. Most suspects, excluding illegal aliens, are expected to have a SSN. Therefore, we only looked into SSN records in our sample data and found 58.3% of the suspects used a falsified SSN. Also, in the falsified SSNs, 96% had no more than two digits different from the corresponding correct ones, for example, "123-45-6789" may be falsified as "123-4**6**-6789" or "123-4**6**-**9**789." We found it rare for criminals to deceive by giving a totally different SSN. In our sample data, only one suspect used a deceptive SSN completely different from his real one.

It is possible for a suspect to forget his or her SSN and unintentionally give an incorrect SSN. It is important to note that giving a false SSN does not automatically flag a deceptive record. It just tells police officers to investigate further. When we compared SSNs between two records, we examined other fields as well. If the SSN was the only altered field in a comparison, the person who reported those two records may have simply forgotten his or her number. This was not considered as deception in our case study and we only considered ID deceptions that were accompanied by deception in other fields.

Threshold	Accuracy	False Negative*	False Positive**
0.4	76.60%	23.40%	0.00%
0.45	92.20%	7.80%	0.00%
0.46	93.50%	6.50%	2.60%
0.47	96.10%	3.90%	2.60%
0.48	**97.40%**	**2.60%**	**2.60%**
0.49	97.40%	2.60%	6.50%
0.5	97.40%	2.60%	11.70%

* False negative: consider dissimilar records as similar ones
** False positive: consider similar records as dissimilar ones

Table 2. Accuracy comparison based on different threshold values.

Residency deception. Suspects usually made changes to only one portion of the full address; street numbers and street types were typically altered. In our sample, 33.3% of the criminals had changed one portion of the address. Deception in more than one component of the address was not found.

Threshold	Accuracy	False Negative	False Positive
0.48	94.0%	6.0%	0.0%

Table 3. The accuracy of linkage in the testing data set.

Deception Detection Algorithm Design and Experimental Results

To detect the deceptions identified in the taxonomy, we chose the four most significant fields (name, DOB, SSN, and address) for our analysis. The idea was to compare each corresponding field of every pair of records. Disagreement values for each field were summed up to represent an overall disagreement value between two records.

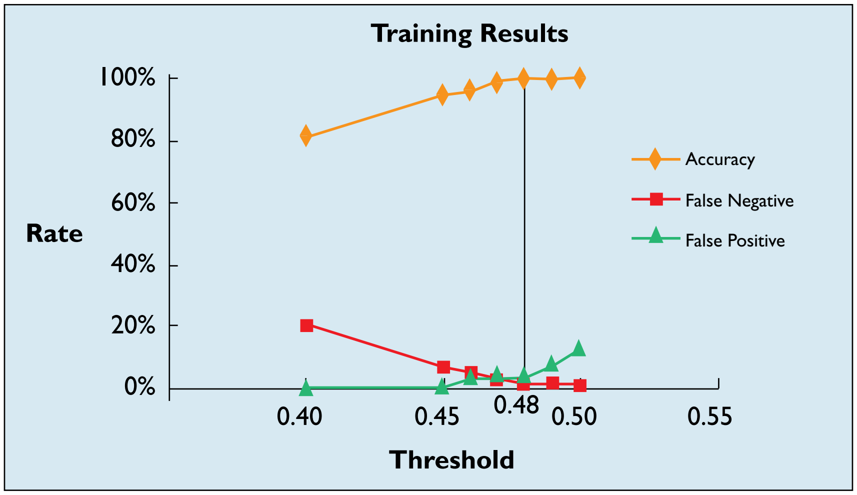

Figure 2. Training accuracy comparison based on different threshold values.

As previously discussed, we used a combination of edit distance and Soundex string comparators. To detect both spelling and phonetic variations between two name strings, edit distance and Soundex disagreement values were computed separately. In order to capture name exchange deception, disagreement values were also computed based on different sequences of first name and last name. We took the disagreement value from the sequence that had the least difference (the minimum disagreement value) between two names. Edit distance itself was used to compare nonphonetic fields of DOB, SSN, and address. Each disagreement value normalized between 0 and 1. The disagreement value over all four fields was calculated by a normalized Euclidean distance function. According to our expert police detective, each field may have equal importance for identifying a suspect. Therefore, we started by assigning equal weights to each field.

Experiment data collection. In order to test the feasibility of our algorithm, a sample set of data records with identified deception was chosen from the police database. At the time, we were not considering records with missing fields. Therefore, we drew from police profiles another set of 120 deceptive criminal identity records with complete information in the four fields. Our veteran Tucson police detective verified that all the records had deception information. The 120 records involved 44 criminals, each of whom had an average of three records in the sample set. Some data was used to train and test our algorithm so that records pointing to the same suspect could be associated with each other.

Training and testing were validated by a standard hold-out sampling method. Of the 120 records in the test bed, 80 were used for training the algorithm, while the remaining 40 were used for testing purposes.

Training results. A disagreement matrix was built based upon the disagreement value between each pair of records. Using the disagreement values in the matrix, threshold values were tested to distinguish between the in-agreement pairs of records and the disagreement pairs. Accuracy rates for correctly recognizing agreeing pairs of records using different threshold values are shown in Table 2. When the threshold value was set to 0.48, our algorithm achieved its highest accuracy of 97.4%, with relatively small false negative and false positive rates, both of which were 2.6% (see Figure 2).

Testing results. Similarly, a disagreement matrix was built for the 40 testing records by comparing every pair of records. By applying the optimal threshold value 0.48 to the testing disagreement matrix, records having a disagreement value of less than 0.48 were considered to be pointing to the same suspect and were associated together. The accuracy of linkage in the testing data set is shown in Table 3. The result shows the algorithm is effective (with an accuracy level of 94%) in linking deceptive records pointing to the same suspect.

Conclusion

We have presented a record-linkage method based on string comparators to associate different deceptive criminal identity records. The experimental results have shown the method to be promising. The testing results also show that no false positive errors (recognizing related records as unrelated suspects) occurred, which means the algorithm has captured all deceptive cases. On the other hand, all the errors occurred in the false negative category, in which unrelated suspects were recognized as being related. In that case, different people could mistakenly be considered the same suspect. This might be caused by the overall threshold value gained from the training process. The threshold value was set to capture as many true similar records as possible, nonetheless, a few marginal dissimilar pairs of records were counted as being similar. Currently, an investigator-guided verification process is needed to alleviate such a problem. An adaptive threshold might be more desirable for making an automated process in future research.

The proposed automated deception detection system will also be incorporated into the ongoing COPLINK project [2] under development since 1997 at the University of Arizona's Artificial Intelligence Lab, in collaboration with the TPD and the Phoenix Police Department (PPD). It continues to be funded by the National Science Foundation's Digital Government Program. C

References

1. Burgoon, J.K., Buller, D.B., Guerrero, L.K., Afifi, W., and Feldman, C. Interpersonal deception: XII. Information management dimensions underlying deceptive and truthful messages. *Communication Monographs 63* (1996). 50–69.
2. Hauck, R.V., Atabakhsh, H., Ongvasith, P., Gupta, H., and Chen, H. Using COPLINK to analyze criminal-justice data. *IEEE Computer* (Mar. 2002).
3. Jaro, M.A. *UNIMATCH: A Record Linkage System: User's Manual*. Technical Report, U.S. Bureau of the Census, Washington, DC, 1976.
4. Levenshtein, V.L. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*, (1966), 707–710.
5. Newcombe, H.B., et al. Automatic linkage of vital records. *Science 130*, 3381 (1959), 954–959.
6. Porter, E.H. and Winkler, W.E. Approximate string comparison and its effect on an advanced record linkage system. *Record Linkage Techniques* (1997), 190–202.
7. Vrij, A. *Detecting Lies and Deceit: The Psychology of Lying and the Implication for Professional Practice.* John Wiley, 2000.
8. Winkler, W.E.. The state of record linkage and current research problems. In *Proceedings of the Section on Survey Methods of the Statistical Society of Canada*,1999. (Also in technical report, RR99/04. U.S. Census Bureau; www.census.gov/srd/papers/pdf/rr99-04.pdf.)

Gang Wang (gang@bpa.arizona.edu) is a doctoral student in the Department of Management Information Systems, The University of Arizona, Tuscon.

Hsinchun Chen (hchen@bpa.arisona.edu) is McClelland Endowed Professor in the Department of Management Information Systems, The University of Arizona, Tuscon.

Homa Atabakhsh (homa@bpa.arizona.edu) is a principal research specialist in the Department of Management Information Systems, The University of Arizona, Tuscon.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

© 2004 ACM 0002-0782/04/0300 $5.00