

Discovering Identity Problems: A Case Study

G. Alan Wang¹, Homa Atabakhsh¹, Tim Petersen², Hsinchun Chen¹

¹ Department of Management Information Systems, University of Arizona, Tucson, AZ 85721
{gang, homa, hchen}@eller.arizona.edu

² Tucson Police Department, Tucson, AZ 85701
Tim.Petersen@tucsonaz.gov

Abstract. Identity resolution is central to fighting against crime and terrorist activities in various ways. Current information systems and technologies deployed in law enforcement agencies are neither adequate nor effective for identity resolution. In this research we conducted a case study in a local police department on problems that produce difficulties in retrieving identity information. We found that more than half (55.5%) of the suspects had either a deceptive or an erroneous counterpart existing in the police system. About 30% of the suspects had used a false identity (i.e., intentional deception), while 42% had records alike due to various types of unintentional errors. We built a taxonomy of identity problems based on our findings.

1. Introduction

Identity resolution is central to fighting against crime and terrorist activities in various ways. Identity information, in many cases, is unreliable due to intentional deception [7] or data entry errors. Commercial database systems search records mainly based on exact-matches. Records that have very minor changes may not be returned by searching on an exact-match. This causes problems for information retrieval which in law enforcement and intelligence investigations would cause severe consequences [3].

In this research we look into the problems that produce difficulties in retrieving identities. The paper is organized as follows: In section 2 we briefly introduce the definition of personal identity based on related research. In section 3 we describe a case study conducted in a local law enforcement agency. Results are analyzed and summarized to create a taxonomy of identity problems. In section 4 we summarize our findings and suggest techniques that improve identity information retrieval.

2. Background

An identity is a set of characteristic elements that distinguish a person from others [2, 5]. Since identity theft/fraud has become a serious problem, some research has been done specifically on identity issues.

In their report on identity fraud, the United Kingdom Home office [4] identified three basic identity components: attributed identity, biometric identity, and

biographical identity. Clarke's identity model [1] gives a more detailed classification of identity information. He argues that identity information falls into one of the five categories: social behavior, names, codes, knowledge, tokens, and biometrics. These works mainly focus on the representation of identities, addressing what it takes to distinguish a person from others. However in the real world, it is impossible to collect all types of identity information for every person. And information collected in the real world is far from perfect. For example, criminal suspects might intentionally use false identities when being confronted by police officers. Those problems make identity identification a difficult job.

3. A Case Study on Identity Problems

A rich source for research into identity problems is the records management systems of local police departments. We chose Tucson Police Department (TPD) as our test bed. TPD serves a relatively large population that ranks 30th among US cities with populations of over 100,000, and Tucson's crime index ranked around 20th highest among US metropolitan areas. We hope that the results of the case study conducted at the TPD can be generalized to other law enforcement agencies.

An identity record in the TPD system consists of many attributes such as name, DOB (date of birth), ID numbers (e.g., SSN, Driver's License Number), gender, race, weight, height, address, and phone number. Biometrics attributes such as fingerprints are not available to this study due to privacy and security reasons. An identity record may not have values in all of its attributes. Name is a mandatory attribute and always has a value. Other attribute values are allowed to be empty or are assigned a default value when not available (e.g., the default value for height in the TPD is 1).

3.1 Data Collection

TPD has 2.4 million identity records for 1.3 million people. Some people may have more than one identity record associated with them. We suspect there might be deception and errors in duplicate records. We first collected identity records for people who had more than one identity in the TPD database. To our disappointment, we found that multiple identities associated to the same person were exact duplicates. Most of them had exactly the same values in such attributes as name, DOB, ID numbers, etc. According to a TPD detective, two identities were associated only when police investigation happened to catch the duplicates. However, duplicate records that differ too much might be less noticed than exact duplicates during police investigation. Therefore, they might not have been associated in the database.

We then randomly drew 200 unique identity records from the TPD database. We considered them to be a list of "suspects" that we were trying to find any matching identities for in the TPD database. Given the huge amount of identity records in the TPD, it is nearly impossible to manually examine every one of them. We used an automated technique [7] that computes a similarity score between a pair of identities. This technique examines only the attributes of name, address, DOB, and SSN. It first measures the similarity between values in each corresponding attribute of the two

identities and then calculates an overall similarity score as an equally weighted sum of the attribute similarity measures. We used this technique to compare each suspect's identity to all other identities in the database. For each suspect's identity, we chose the 10 identity records that had the highest similarity scores.

We verified the 10 possible matches for each of the 200 suspects by manually examining their key attributes. Each possible match was classified into one of the four categories defined in Table 1. The first two categories, D and E, imply a true match. A matching identity was considered to be an error when identical values were found in key attributes such as name and ID numbers. A matching identity was considered deceptive when key attribute values such as name, DOB and ID numbers were not identical but showed similar patterns. If a matching identity had very different values on those key attributes, it was considered a non-match. If a matching identity was not in any aforementioned category, it was categorized as U (uncertain).

Table 1. Categories into which possible matches are classified

Category	Description
D	Intentional <u>D</u> eception
E	Unintentional <u>E</u> rrors
N	<u>N</u> on-match
U	<u>U</u> ncertain (too little information to make a call)

3.2 Preliminary Evaluation

We asked a TPD detective who has worked in law enforcement for 30 years to verify our categorization results. He agreed on most of our decisions. It surprised us that more than half (55.5%) of the suspects had either a deceptive or an erroneous counterpart existing in the TPD system. About 30% of the suspects had used a false identity (i.e., intentional deception), while 42% had records alike due to various types of unintentional errors. As the numbers imply, some suspects may have both deceptive and erroneous records in the TPD system.

3.3 A Taxonomy of Identity Problems

As shown in Figure 1, we built a taxonomy of identity problems based on our findings. Among others available in the TPD, attributes such as name, DOB, ID numbers, and address indicate deception or errors in most cases. Attributes such as weight, height and race are usually estimated through visual inspection by police officers, and their values are by nature indefinite. We do not consider value differences for these attributes deception or errors. Although gender is also visually inspected and is often definite, we did not find any evidence in our sampled data that officers were deceived or made a mistake on that.

Deceptive identities and erroneous identities exhibit quite different characteristics. We discuss the two types of identity problems respectively in the rest of this section.

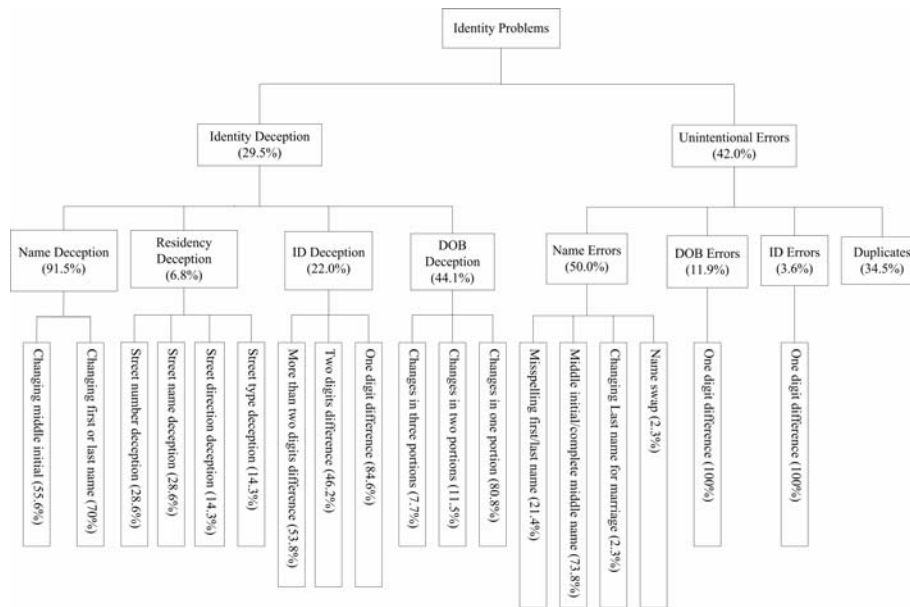


Fig. 1. A taxonomy of identity problems

Identity Errors. Erroneous identities were found to have discrepancy in only one attribute, or having no discrepancy at all (i.e., duplicates). There were 65.5% of erroneous identities that had slightly altered values in either name (50.0%) or DOB (11.9%) or ID numbers (3.6%). They had errors in only one attribute and had other attribute values identical to those of the corresponding true identity. The rest of the erroneous identities (34.5%) were merely duplicates. Their attribute values were all identical to those of the corresponding true identity.

Values in name errors were altered in different ways. 73.8% of them either had their middle names missing or switched between using a middle initial and using a full middle name. Misspelling was found in 21.4% of name errors. There was only one instance where first name and middle name were swapped (2.3%). There was also another instance in which the name was altered by having a different last name after marriage (2.3%). Certainly, such an instance only applies to female identities.

A DOB value consists of three portions: a year, a month, and a day, e.g., 19730309. In our study most of the erroneous DOB values were only one-digit different from the corresponding true values. It is likely caused by mistyping or misreading from a police report. An erroneous DOB is not necessarily close to its corresponding true value in time. For example, one DOB was altered as “19351209,” while its true value was “19551209.” In this case one-digit difference made 20 years’ difference in time.

Similar to DOB errors, most of the ID errors were also found to be different from the corresponding true values by only one digit.

Identity Deception. Deceptive Identities usually involves changing values in more than one attribute. The value changes are more drastic than those in erroneous identities. Name was found to be the attribute most often subject to deception (91.5%). Less than half of the deceptive identities (44.1%) had altered DOB values and 22% of them had altered ID numbers. There were also 6.8% of the deceptive identities with an altered residential address. We discuss each type of identity deception in details below.

Seventy percent of the deceptive names were altered by changing either first or last name, but not both of them. Also the names were not altered randomly. Popular name changing techniques we found include: using a name that is phonetically similar (e.g., Wendy being altered to Windy), using a nick name (e.g., Patricia being changed to Trish), using a name translated from other languages (e.g., a Spanish name Juan being changed to Johnny), or using someone else's name (e.g., brother's or sister's name). Changing first or last name was often accompanied by changing the middle name.

As opposed to DOB errors, deceptive DOB values can be made by changing more than one portion. We found that 7.7% of deceptive DOBs made changes in all three portions, 11.5% of them made changes in two portions, and 80.8% made changes in one portion. Deceptive DOB values were often altered in a way different from DOB errors. An erroneous DOB value often results in one-digit difference by, for example, mistyping a visually similar number (e.g., typing 0 for 8). One the contrary, a deceptive DOB value may have more than one-digit difference. For example, one subject made himself seem younger by reporting his DOB as "19630506," while his true DOB was "19580506." In deceptive DOB, although there is still only one altered portion, the altered value has a two-digit difference from the true value. There was also evidence that people alter DOB values using some patterns. One subject reported his DOB as "19730203," while his true DOB was "19730102." One can notice the pattern of adding one to both month and day values. Switching digits is another popular technique. For example, one subject reported "19380129," while his true DOB was "19390128."

Changes made in deceptive ID numbers were also more drastic than ID errors. 53.8% of deceptive ID numbers were altered more than three digits. 46.2% of them were altered by more than two. Making a one-digit change was also common in ID deception (84.6%). Those percentages do not sum up to one hundred percent because an identity record may have multiple ID numbers associated. One who deceives on one of his/her ID numbers will be very likely to deceive on other numbers associated with him/her. Some techniques used for DOB deception, such as switching digits, were also found in ID number deception.

Residential address is unreliable in determining a person's identity because many people move frequently. We found that 6.8% of the suspects deceived at least once on address information. An address consists of four main components: street number, street direction, street number, and street type. Sometimes there is also a suite/apartment number. 28.6% of address deception altered street numbers or street names, while 14.3% of address deception altered street direction or street type. It seems people are more used to altering textual values such as numbers and names than to changing nominal values such as street directions and types.

4. Conclusion and Discussion

In this study we examine the identity problems that result in difficulties for identity resolution. Two types of problems, including intentional deception and unintentional errors, were found in real law enforcement records. Attribute values were altered more drastically in deception than in errors. In most of the cases, altered values looked very similar to the corresponding true values.

Our future work is to develop an automated identity resolution technique that takes our findings on identity problems into account. Techniques that improve identity information retrieval should locate identity information in an approximate rather than exact manner. Similarity measures need to be defined for different attributes based on the characteristics presented by both deceptive and erroneous values. A string comparator such as Edit Distance [6] may be a good candidate in some cases. But in some other cases, the similarity measure can be more complex. For example, a DOB similarity measure needs to capture the pattern of adding one to both month and day values. A decision model is necessary to determine whether an identity should be retrieved given a set of similarity measures on attributes. Finally, the techniques have to be automated because the huge amount of data prohibits any possibility of manual operations.

Acknowledgement

This project has been primarily funded by the following grant: NSF, Digital Government Program, "COPLINK Center: Social Network Analysis and Identity Deception Detection for Law Enforcement and Homeland Security," #0429364, 2004-2006.

References

1. Clarke, R.: Human Identification in Information Systems: Management Challenges and Public Policy Issues. *Information Technology & People* 7, 4 (1994) 6-37
2. Donath, J. S.: Identity and Deception in the Virtual Community. In: M. a. K. Smith, P. (ed.): *Communities in Cyberspace*. Routledge, London, (1998)
3. GAO: Law Enforcement: Information on Timeliness of Criminal Fingerprint Submissions to the FBI. GAO-04-260, United States General Accounting Office (GAO) (2004)
4. HomeOffice, U. K.: Identity Fraud: A Study. United Kingdom HomeOffice (2002)
5. Jain, A. K., Prabhakar, S., Hong, L., and Pankanti, S.: FingerCode: A fingerbank for fingerprint presentation and matching. In: *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (1999)
6. Levenshtein, V. L.: Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady* 10, (1966) 707-710
7. Wang, G., Chen, H., and Atabakhsh, H.: Automatically Detecting Deceptive Criminal Identities. *Communications of the ACM* 47, 3 (2004) 71-76