Burglary Data Mining - A Three Tiered Approach: Local, State, And Nation-Wide

Anna Koufakou[1], Abhijit Wakchaure[1], Olcay Kursun[2], Michael Georgiopoulos[1], Kenneth Reynolds[3], Ronald Eaglin[2]

[1]Department of Electrical and Computer Engineering
University of Central Florida, Orlando, FL 32816
Phone: (407)-823-5338 Fax: (407)-823-5835
{akoufako@ucf.edu, awakchau@ucf.edu, michaelg@mail.ucf.edu}

[2]Department of Engineering Technology
University of Central Florida, Orlando, FL 32816
Phone: (407)-823-5937 Fax: (407)-823-5483
{okursun@mail.ucf.edu, reaglin@mail.ucf.edu}

[3]Department of Criminal Justice and Legal Studies
University of Central Florida, Orlando, FL 32816
Phone: (407)-823-2943 Fax: (407)-823-5360
kreynold@mail.ucf.edu

**Presented at TU GIS 2005 – The 2nd Annual GIS Symposium at TU,
Troy University, Troy, Alabama USA
May 10-11th 2005**

ABSTRACT

The obvious need for using modern computer networking capabilities to enable the effective sharing of information among jurisdictions has resulted in data-sharing systems which produce huge amounts of data. The vast data thus made available needs to be effectively utilized and analyzed in the form of a data-mining tool for achieving law enforcement goals. In Hillsborough County, using Florida's data sharing network, detectives have arrested a suspect and cleared forty residential burglaries. The suspect have told the detectives that he would steal in one county and sell the stolen property in another county believing it would be less likely he would be caught. In this work, we focus on burglary-related data obtained from various law enforcement agencies, and develop an artificial intelligence (AI) tool to identify potential suspects for unsolved burglaries and to predict their associations with those burglaries. In the process of building such a tool, there are various major issues to be addressed. One major problem is dirty data that we have focused in this work.

Keywords and phrases: alias finding, data mining, data sharing, dirty data, duplicate elimination, fuzzy name matching, pattern discovery, soundex.

1.0. INTRODUCTION

Our ultimate goal in this research field is to mine burglary-related data obtained from various law enforcement agencies, and develop an artificial intelligence (AI) tool to identify potential suspects for unsolved burglaries and to predict their associations with those burglaries. One big challenge in such data mining applications is to eliminate duplicate records and identify aliases. In this paper, we will focus particularly on this issue of data quality considerations. As with any sort of data analysis, the data to be analyzed needs to be thoroughly checked for any errors, mistyped words or such similar considerations. The data, thus, needs to be clean and ready for any type of analysis and interpretation.

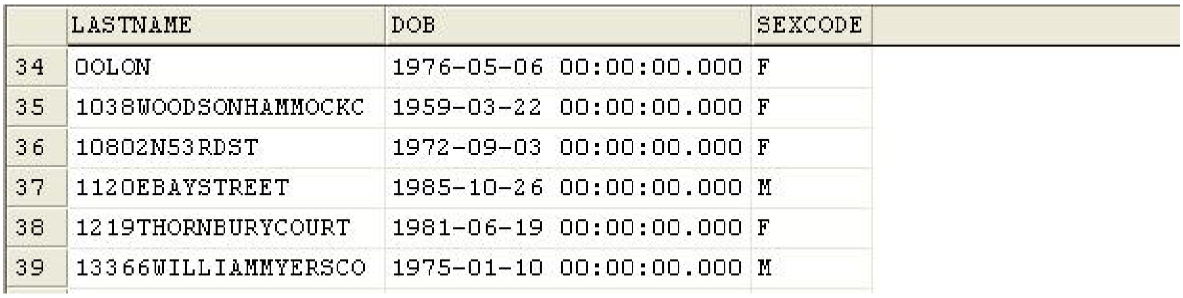
In Section 2, we describe the criminal dataset we have and exemplify some of the common problems with data misentries. In Section 3, we review existing methods for string comparison. We present our algorithm for fuzzy record matching in Section 4. Section 5 is on the status of getting data from multiple jurisdictions. We conclude in Section 6.

2.0. DATA QUALITY CONSIDERATIONS

In our study, we have a very large database containing almost half a million records. With the increasing number of records, the chances of having 'dirty data' (aliases, misentries, etc...) increases. The database contains two broad schemas, one incorporating information pertaining to persons and vehicles, and the other focusing on information about pawned item details. The two schemas altogether contain about 20-25 tables. One of the most important table is the 'Report' table, which is referenced by a lot of tables and contains detailed information about each report which is generated. Also, we have the 'Person_name' and 'Person' tables, which contain important information about the persons (suspects, victims, witnesses, etc...). Another important table is the 'Ticket_item' table which holds information about the details of pawned items.

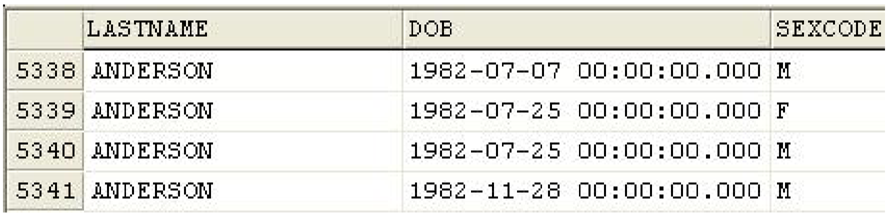
We did a preliminary data exploration to find out how clean the data really was. A person's details typically include the person's last name, first name, middle name or middle initial, sex, race, Date of Birth (DOB), etc... In our analysis, we used the 'lastname', 'firstname', 'sex', 'race', and DOB' as our attributes for comparison as siggested by Wang (*et al.* 2004).

Following are some of the examples which highlight some of the discrepancies in the data:



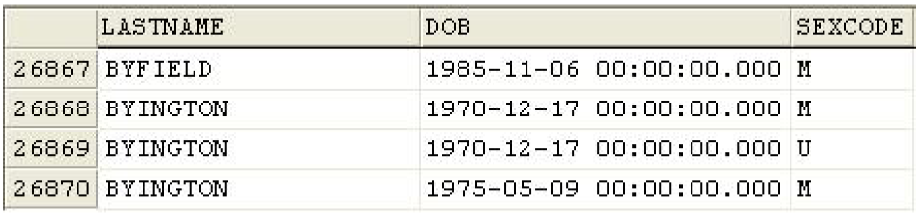
	LASTNAME	DOB	SEXCODE
34	OOLON	1976-05-06 00:00:00.000	F
35	1038WOODSONHAMMOCKC	1959-03-22 00:00:00.000	F
36	10802N53RDST	1972-09-03 00:00:00.000	F
37	1120EBAYSTREET	1985-10-26 00:00:00.000	M
38	1219THORNBURYCOURT	1981-06-19 00:00:00.000	F
39	13366WILLIAMMYERSCO	1975-01-10 00:00:00.000	M

As can be seen above, the 'lastname' field has been mistakenly populated with the addresses, and in this case street names have been entered. We also have last names entered as ‘UNKNOWN’ or ‘UNKNOW’, etc. which are examples of typos, misentries, or missing data.



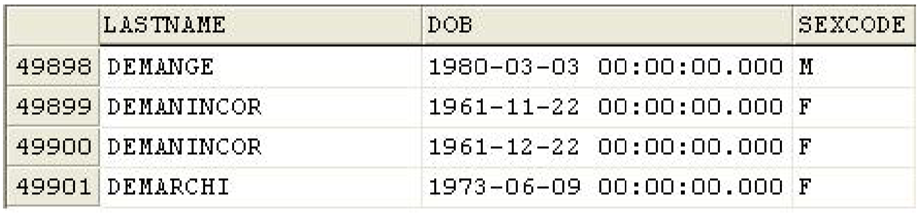
	LASTNAME	DOB	SEXCODE
5338	ANDERSON	1982-07-07 00:00:00.000	M
5339	ANDERSON	1982-07-25 00:00:00.000	F
5340	ANDERSON	1982-07-25 00:00:00.000	M
5341	ANDERSON	1982-11-28 00:00:00.000	M

We see above that two records may have the same lastname and the same DOB, however the sexcode may be 'M' in one case and 'F' in the other, and thus we are not sure if these are the same individual and what is the right sexcode.



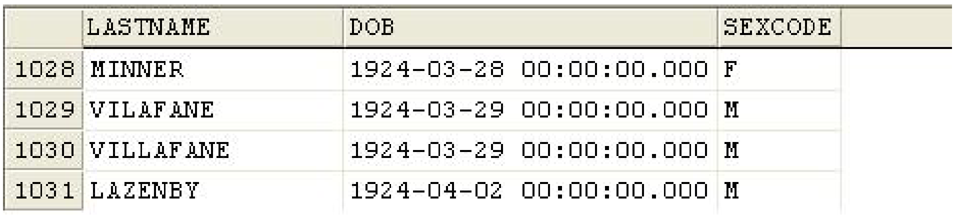
	LASTNAME	DOB	SEXCODE
26867	BYFIELD	1985-11-06 00:00:00.000	M
26868	BYINGTON	1970-12-17 00:00:00.000	M
26869	BYINGTON	1970-12-17 00:00:00.000	U
26870	BYINGTON	1975-05-09 00:00:00.000	M

Again above, we have the same 'lastname' and 'DOB', however the sexcode entered is 'M' and also as 'U'(meaning Unknown).



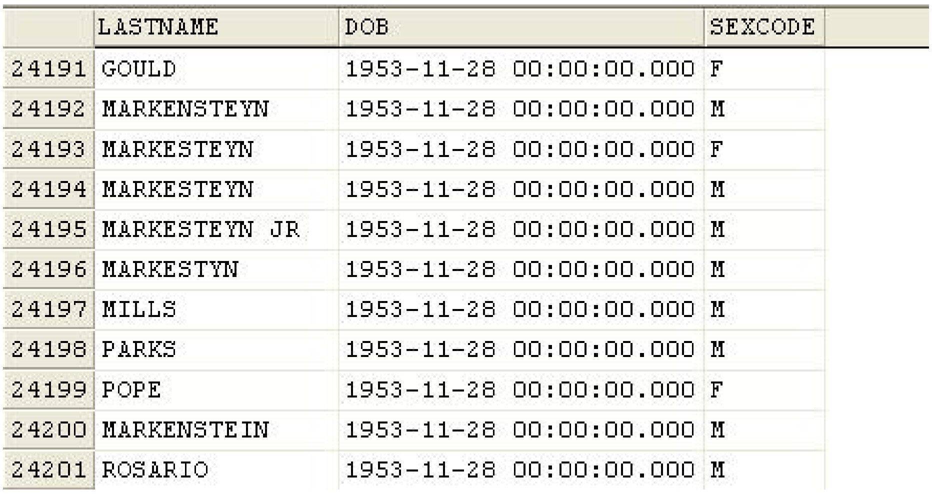
	LASTNAME	DOB	SEXCODE
49898	DEMANGE	1980-03-03 00:00:00.000	M
49899	DEMANINCOR	1961-11-22 00:00:00.000	F
49900	DEMANINCOR	1961-12-22 00:00:00.000	F
49901	DEMARCHI	1973-06-09 00:00:00.000	F

Are these two Demanincor's different individuals or is it a DOB error? Since it differs only by one digit of DOB, it can be taken as the same individual with a certain probability. We will explain this approach in Section 5.



	LASTNAME	DOB	SEXCODE
1028	MINNER	1924-03-28 00:00:00.000	F
1029	VILAFANE	1924-03-29 00:00:00.000	M
1030	VILLAFANE	1924-03-29 00:00:00.000	M
1031	LAZENBY	1924-04-02 00:00:00.000	M

Are Vilafane and Villafane two different individuals or is there a small typo in the lastname? In Section 5, we will use edit distance to measure how different two strings are in order to conclude they are the same individual with certain probability.



	LASTNAME	DOB	SEXCODE
24191	GOULD	1953-11-28 00:00:00.000	F
24192	MARKENSTEYN	1953-11-28 00:00:00.000	M
24193	MARKESTEYN	1953-11-28 00:00:00.000	F
24194	MARKESTEYN	1953-11-28 00:00:00.000	M
24195	MARKESTEYN JR	1953-11-28 00:00:00.000	M
24196	MARKESTYN	1953-11-28 00:00:00.000	M
24197	MILLS	1953-11-28 00:00:00.000	M
24198	PARKS	1953-11-28 00:00:00.000	M
24199	POPE	1953-11-28 00:00:00.000	F
24200	MARKENSTEIN	1953-11-28 00:00:00.000	M
24201	ROSARIO	1953-11-28 00:00:00.000	M

As can be seen from the above example, there are quite some variations for the same lastname, and sometimes the sexcode can be entered incorrectly, making it difficult to identify the individual's correct information (if it is the same individual afterall).

Thus, we have shown that there are quite a number of data quality considerations, which we need to account for while we carry out data analysis, without which we shall be unable to interpret the results correctly.

3.0. RELATED RESEARCH

The problem of matching differently spelled names located in numerous records in one or more different files is closely related to Record Linkage. This is defined by Winkler (1999) as the method of finding duplicate records in a file or matching different records in different files. Newcombe (*et al.* 1959) was to first to design a computerized approach to record linkage with his study in matching a marriage record in a marriage file system with a birth record in a birth profile system. The latest record linkage techniques encompass elements from an array of different areas, such as computer science, and operations research.

The main idea behind all these techniques is comparing two or more strings in order to decide if they both represent the same string, or the same individual in our case. In this case, a *string comparator* must be developed to establish the similarity between different attributes, such as names, date of birth, etc. In addition, different weights must be applied to reflect the importance of the different attributes in the matching of the individual records. For example, a matching Last Name might be given more importance and thus a higher weight compared to matching Address information.

The main string comparators found in the literature are phonetic and spelling based. Newcombe (*et al.* 1959) used the Russel Soundex Code to encode last names. Soundex is used to represent phonetic patterns in a word, by encoding it as the first letter followed by a three-digit number. This way, differently spelled names that are pronounced similar will have the same Soundex code, e.g. "Pierce" and "Pearse" are both coded as "P620". Although Soundex is very successful in contrast to its simplicity, often it produces false results, e.g., "Christie":(C623) and "Kristie":(K623) are pronounced similarly, but have different Soundex encodings. Also, "Kristie" and "Kirkwood" share the same Soundex code but are very different names.

In contrast, spelling string comparators check the spelling differences between strings instead of phonetic encodings. Specifically, Jaro (1976) presented a spelling string comparator which checks for typographical errors, mainly concentrating on inserting, deleting, and transposing characters in a string. Another method used to compare strings is measure their "edit distance" defined by Levenshtein (1966). This can be viewed as the minimum number of single character that need to be inserted into, deleted from, and/or substituted in one string to get another. This measure outperforms Jaro's because it can handle different kind of string patterns. However, these techniques are more complicated, which results in higher space and time complexities in run-time.

Similar work to ours has been described in (Wang *et al.* 2004), which focuses in identifying deceptive criminal identities, i.e. in matching different records that correspond to the same individual mainly due to false information provided by these individuals. According to the authors in (Wang *et al.* 2004), the main attributes or fields needed for this work are Name, DOB, SSN, and Address. Also, career criminals tend to only use partially deceptive names most of the time, change only one portion of their residency information, or of their date of birth, or of their identification number (ID number or Social Security Number). However, the results of this work were based on clean data in small numbers, while our work will concentrate in finding matching records for individuals regardless of missing information and dirtiness of data as described in the next section.

4.0. OUR APPROACH AND PRELIMINARY RESULTS

In order to detect partial matching identities in querying the database, we have decided to use the following fields with their weights (in per cent) shown in parentheses: Lastname (40%), Firstname (20%), Sex (10%), Race (10%), DOB (20%), in total summing up to 100%. These weights are adjusted according to what information is available for use. In other words, if the user does not know (or enter) the first name for the search, then its weight is shifted to other fields in proportion to their relative importance. Together with the query parameters, the user enters a threshold value, which controls requested strength of a match between the query parameters and the available records in the database. We have used Levenshtein's (1966) edit distance for determining the match for the first and last names. In order to convert the edit distance to a match score, we have used the formula given in Eq. 1. Using Eq. 2, we obtained final match score, which is compared against the threshold.

$$match(s_f, r_f) = 1 - \frac{LD(s_f, r_f)}{\max(length(s_f), length(r_f))}, \qquad (1)$$

where s_f represents the field f of the search parameters, r_f represents the field f of record r, and LD represents Levenshtein's edit distance.

$$match\_score = \sum_f match(s_f, r_f) * weight(f), \qquad (2)$$

where f represents all possible fields that are not blank (i.e. specified by the user as query parameter and also not blank in the record that is currently being analyzed).

A record is returned in query's output only if match score is greater than or equal to the threshold. Clearly, setting the threshold to 100% means no fuzzy matches will be returned:

QUERY

Last Name?	BALDWIN	First Name?	DANNY
Sex?	M	Race?	W
DOB?	12/31/71	Threshold %?	100%

RETURNS (LASTNAME, FIRSTNAME, SEX, RACE, DOB, %MATCH)

BALDWIN, DANNY, M, W, 12/31/71, 100.0%

However, if we enter a threshold of 85%, then the same query runs as follows, returning two records. Note that the race information was probably entered wrong as B instead of M, which reduces the match score to a still acceptable 90%.

QUERY

Last Name?	BALDWIN	First Name?	DANNY
Sex?	M	Race?	W
DOB?	12/31/71	Threshold %?	85%

RETURNS (LASTNAME, FIRSTNAME, SEX, RACE, DOB, %MATCH)

BALDWIN, DANNY, M, **W,** 12/31/71, **100.0%**

BALDWIN, DANNY, M, **B,** 12/31/71, **90.0%**

To demonstrate the meaning of shifting the weights according to availability of data fields, suppose we run the same query above without entering DOB. The confidence of 90% (for the second record) reduces to 87.5%. If DOB is entered as in the query above, the match between the entered DOB and the DOB of the record gives extra confidence. If DOB is not entered as in the query below, its weight shifts towards the other fields, thus, increasing the weight of race field as well. Therefore, if there is a mismatch of races between the query parameters and a record, that results in lower match score for that record:

QUERY

Last Name?	BALDWIN	First Name?	DANNY
Sex?	M	Race?	W
DOB?	**UNKNOWN**	Threshold %?	85%

RETURNS (LASTNAME, FIRSTNAME, SEX, RACE, DOB, %MATCH)

BALDWIN, DANNY, M, W, 12/31/71, 100.0%

BALDWIN, DANNY, M, B, 12/31/71, **87.5%**

Some other exemplary queries are shown below:

QUERY

Last Name?	ABOODI	First Name?	UNKNOWN
Sex?	UNKNOWN	Race?	UNKNOWN
DOB?	UNKNOWN	Threshold %?	85%

RETURNS (LASTNAME, FIRSTNAME, SEX, RACE, DOB, %MATCH)
ABOODI, ARASH, M, W, 12/26/**62**, 100.0%
ABOODI, ARASH, M, W, 12/26/**82**, 100.0%

QUERY

Last Name?	BARKLEY	First Name?	JOANN
Sex?	F	Race?	UNKNOWN
DOB?	UNKNOWN	Threshold %?	95%

RETURNS (LASTNAME, FIRSTNAME, SEX, RACE, DOB, %MATCH)
BARKLEY, **JOANN,** F, B, 12/03/82, **100.0%**
BARKLEY, **JOANNA,** F, B, 12/03/82, **95.2%**

5.0. DATA SHARING FOR DETECTING MULTI-JURISDICTIONAL CRIMES

The extraction of knowledge (knowledge mining) from massive amounts of data has been the focus of many research papers in the recent years. The huge size and complexity of the data renders several existing algorithms unacceptable although they might have been proven successful on smaller data sets. These traditional data mining algorithms typically assume that the data can reside in memory, which can no longer be accepted. Thus the additional challenge originates from the fact that today's data are distributed or dispersed over a wide geographic region which have to be integrated and combined so that they can be considered as a whole. Many issues arise from the distributed nature of the data, such as maintaining security, minimizing communication delays, load balancing, securing data integrity and autonomy, etc.

In our case, the methods and techniques that will be developed will be part of the Law Enforcement Data Sharing Consortium which is already in place as a data sharing system formed to allow public safety organizations to exchange appropriate information in an efficient and economic way in Florida.

Data Sharing Consortium is a group of law enforcement agencies who have decided to design, develop, fund, and implement a data sharing infrastructure on their own. These agencies are not waiting for some state or federal agency to solve this problem for them. They are not willing to spend large amounts of money. They have formed a public partnership with each other and with the University of Central Florida (UCF), and the group has worked together in this non-profit configuration since August 2000 to implement such a system. Already, more than 80 of the state's 355 police agencies have begun sharing internal arrest records and other information that will reduce much of their manual labor and phone calls. By midyear, the project's founders hope to have at least 100 agencies connected to a network, with the rest coming online by 2007. The system, which relies on participant fees and is not affiliated with any private company, lets agencies tap into pawn-shop records, sex-offender data and field-information reports. The latter are completed by officers who interview or detain people without arresting them (Gutierrez 2005).

The interoperability that this system offers supports traditional crime suppression objectives and is crucial in this era of heightened domestic security. This system will allow access to an unmatched amount of information that was previously inaccessible. It can be utilized by every member of the agency. This information provides an opportunity for agencies to address crime control issues that cross jurisdictional boundaries. It can also save countless man-hours by allowing agency personnel to query a system to obtain information that they otherwise would attempt to obtain by making numerous and time consuming phone calls.

Detectives state that in today's society, crooks do not stay in the same place, they are very mobile. It is not uncommon that a detective in one jurisdiction looking for a suspect finds out later that the guy he wanted was sitting in a county jail two counties away, but they did not know it and the suspect was released. Detectives can now go to one place in order to learn something about a suspect, where he or she has been arrested; whether or not he or she has been in jail; or you need information for an ongoing investigation, they enter the information and within seconds get a thorough answer (Gamble 2004, Pattavina 2005).

Year 2004 proved to be a highly successful and productive year for the Sheriff's Offices and Police Departments connected with Florida's data sharing network. The network has helped snare at least 200 suspects. Armed robbers, burglars, home invaders, sex offenders, and those pawning stolen property were identified and apprehended through data sharing by detectives in Tampa, Orlando, Hillsborough County, Polk County, Orange County, Seminole County, Osceola County, Citrus County, and many other jurisdictions (Saviak 2005).

6.0. FUTURE WORK

The next step in burglary data mining is to develop a tool that applies a pattern matching technique to arrive at a set of possible suspects for a burglary. The tool will also automate the task of visualizing each suspect's timeline (see Figure below) of known criminal activity and also gathers information about shared pawn data incrementally at the local, state, and national level using the queries powered by our string/record matching technique on the data obtained from Data Sharing Consortium. The tool, then, evaluates the strength of a possible match between a suspect's timeline, related pawn records, and the unsolved burglary. The AI tool thus developed will assist the law enforcement personnel in taking advantage of the shared data to tackle burglaries faster, for otherwise laborious or even unsolvable cases.

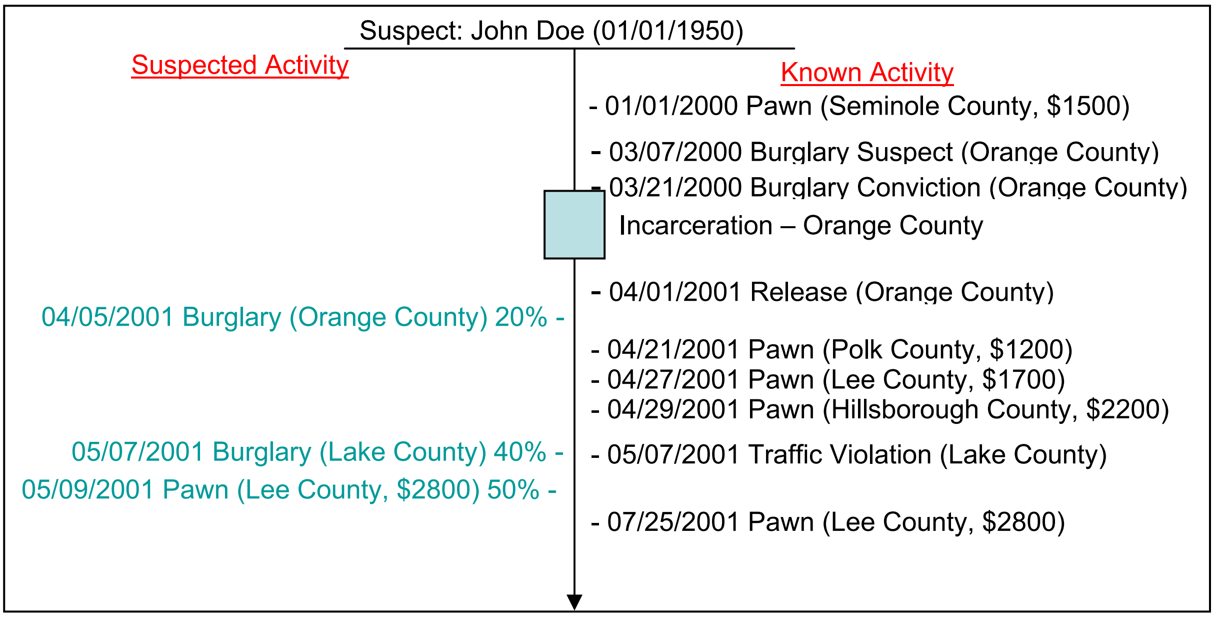

Suspect: John Doe (01/01/1950)
Suspected Activity
Known Activity
- 01/01/2000 Pawn (Seminole County, $1500)
- 03/07/2000 Burglary Suspect (Orange County)
- 03/21/2000 Burglary Conviction (Orange County)
Incarceration – Orange County
- 04/01/2001 Release (Orange County)
04/05/2001 Burglary (Orange County) 20% -
- 04/21/2001 Pawn (Polk County, $1200)
- 04/27/2001 Pawn (Lee County, $1700)
- 04/29/2001 Pawn (Hillsborough County, $2200)
05/07/2001 Burglary (Lake County) 40% -
- 05/07/2001 Traffic Violation (Lake County)
05/09/2001 Pawn (Lee County, $2800) 50% -
- 07/25/2001 Pawn (Lee County, $2800)

Figure. Timeline of John Doe

Since we rely on Data Sharing Consortium to supply the data, the non-centralized data storage approach is vital because it allows agencies to share data without compromising their local control of their data. Specifically, the technology utilized by the Consortium is a distributed system of federated databases which communicate through a secure network using an XML web services communications protocol. Therefore, the aliases found by our approach will need to be stored locally (as alias tables) on the local databases. Upon a query from another agency, the local alias table may be searched for fuzzy matches if the user is interested in partial matches.

ACKNOWLEDGEMENTS

The authors wish to acknowledge helpful discussions with detectives from Orange County Sheriff's Office.

REFERENCES

Gamble, M.R. (2004, July) Data-Sharing System Fights Terrorism. *Converge Online Magazine.*

Gutierrez, P.R. (2005, February 6) Database connects cops. *Orlando Sentinel.*

Jaro, M.A. (1976) UNIMATCH: A Record Linkage System: User's Manual. Technical Report, U.S. Bureau of the Census, Washington, DC.

Levenshtein, V.L. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics*, Doklady, vol. 10, pp. 707–710.

Newcombe, H.B., et al. (1959) Automatic linkage of vital records. *Science,* vol. 3381, pp. 954–959.

Pattavina, A. (2005) Geographic Information Systems and Crime Mapping in Criminal Justice Agencies. In: Pattavina, A. (ed) *Information Technology and the Criminal Justice System*, Sage Publications, Thousand Oaks, CA.

Saviak, J. (2005, January) Data Sharing "Web" Snares Criminals Around The State. *Florida Law Enforcement Data Sharing Consortium Newsletter. http://druid.engr.ucf.edu/datasharing/current_newsletter.doc.*

Wang, G., Chen, H., Atabakhsh, H. (2004) Automatically detecting deceptive criminal identities. *Communications of the ACM*, March 2004, vol. 47(3), pp. 70–76.

Winkler, W.E. (1999) The state of record linkage and current research problems. *Proceedings of the Section on Survey Methods of the Statistical Society of Canada*.