2005

A Multi-Layer Graphical Model for Approximate Identity Matching

G. Alan Wang
University of Arizona, gang@eller.arizona.edu

A Multi-Layer Graphical Model for Approximate Identity Matching

G. Alan Wang
University of Arizona
gang@eller.arizona.edu

ABSTRACT

Many organizations maintain identity information for their customers, vendors, and employees, etc. However, identities being compromised cannot be retrieved effectively. In this paper we first present a case study on identity problems existing in a local police department. The study show that more than half of the sampled suspects have altered identities existing in the police information system due to deception and errors. We build a taxonomy of identity problems based on our findings. The decision to determine matching identities involves some uncertainty because of the problems identified. We propose a probability-based multi-layer graphical model to capture the uncertainty. Experiments show that the proposed model performs significantly better than the searching technique based on exact-match. With 20% of training data labeled, the model with semi-supervised learning achieved performance comparable to that of fully supervised learning.

Keywords

Identity matching, similarity, record linkage, Bayesian network, Graphical model, Semi-supervised learning

INTRODUCTION

Decision-making process often combine different sources of data and knowledge available in various forms (Bolloju, Khalifa and Turban, 2002). One of the knowledge management principles that help achieve collaborative knowledge bases is to provide tools to transform scattered data into meaningful business information and support all types of decision makers (Ba, Lang and Whinston, 1997; Bolloju *et al.*, 2002).

Many organizations maintain identity information for their customers, vendors, and employees, etc. In some cases identity information needs to be carefully managed. For example, a poorly maintained customer database may lead to poor relationship with customers (e.g., a customer would be contacted many times if the database contains duplicate records). There is also a need to match identities among different data sources. For example, when a marketing department has obtained a list of possible customers from other vendors, it is necessary to determine which records already exist in the customer database. It is even more critical to match and integrate identity information in law enforcement and intelligence communities.

Identity information, however, is unreliable in many cases due to intentional deception and unintentional data errors. This causes problems for information retrieval which in law enforcement and intelligence investigations would cause severe consequences. In a recent government report (GAO, 2004), the Federal Bureau of Investigation (FBI) described real cases in which the arresting agencies released suspects from custody when they used false names at the time of the arrest. Later they were found wanted by other jurisdictions. Identity issues have attracted many discussions (Clarke, 1994; HomeOffice, 2002; Kent and Millett, 2002). However, there is no discussion on identity problems that affect effective identity information retrieval. We intend to explore identity problems in this research.

Current techniques deployed in law enforcement and intelligence agencies are neither adequate nor effective for identity matching. Police officers often rely on computer systems to search a suspect's identity against history records in police databases. Generally, computer systems search using exact match queries. Even if a fabricated identity is similar to the true identity recorded in the police database, an exact-match query is unlikely to bring up that record. Techniques that perform inexact identity search have been developed (Badiru, Karasz and Holloway, 1988; Brown and Hagen, 2002; Wang, Chen and

Atabakhsh, 2004). They can detect identities that are similar but not exactly the same. However, these techniques are either not fully automated or reliant on human-generated training datasets. Given the huge amount of identity records kept in police databases, manually generating a training dataset is time-consuming and labor-intensive. We aim to propose an automated technique for identity matching with less human interventions.

The paper is organized in the following order: In the section of literature review we introduce some background information about identity and identity problems. We also review the techniques that are applicable to identity matching. In the next section we report a case study on identity problems using real law enforcement data. In the section of research design we propose a multi-layer graphical model for identity matching. We report the performance of the proposed model in the section of experiments. We summarize our findings and discuss future directions in the last section.

LITERATURE REVIEW

An identity is a set of characteristic elements that distinguish a person from others (Clarke, 1994). There are three basic types of identity components: attributed identity (e.g., name, ID numbers), biometric identity (e.g., fingerprint) and biographical identity (e.g., credit history) (HomeOffice, 2002).

To the best of our knowledge, there are not many literatures focusing on identity problems. The United Kingdom Home Office (HomeOffice, 2002) published a report on identity fraud. They analyzed approaches to establishing false identities and proposed policies for alleviating the problems. However, they did not focus on decision-support techniques that can help to identify false identities. Wang *et al.* (2004) conducted a case study on criminal identity deception. The study showed that criminals preferred half-truth to making up complete lies. Therefore, deceptive identities were found similar to their original ones. However, this study did not consider false identities that were caused by unintentional errors during data entry and manipulation processes. Also, the dataset used in the case study was not randomly selected from the available data pool. Therefore, the conclusions might not represent the complete picture of false identities.

Identity Matching Techniques

Identity matching techniques basically perform automated data association that links suspects to the crime being investigated, ordered from the most possible to the least possible. Each identity record is represented by a vector of attribute values such as name and Date-Of-Birth (DOB). When comparing two identity records, it is natural to first compare the value-pair of each individual attribute. A decision model then maps a vector of comparison features to a binary decision variable: matching or not-matching. There are two types of decision models for identity matching: weighted-sum decision models and probability-based decision models.

A weighted-sum decision model first combines comparison features into a single similarity score upon which the match decision is made. Dey *et al*. (2002) used integer programming to model the entity matching problem. Their approach assumed only one-to-one mapping in two databases. That assumption, however, is not always true in the scenario of identity matching. Brown and Hagen (2002) proposed a similarity-based data association method for associating records of the same suspect or incidents having similar modus operandi. It compares corresponding description attributes of two records and calculates a total similarity measure (TSM) between the two records. Experiments showed that the similarity ratings suggested by the algorithm agreed with those made by human experts. However, this technique only provides a list ordered by ratings and does not differentiate between matches and non-matches. Wang *et al.* (2004) proposed an algorithm specifically designed for detecting deceptive identities. This method makes use of string comparison techniques (i.e., Edit Distance (Levenshtein, 1966)) and searches for inexact matches of suspects' identities in police databases. It examines the attributes of name, address, DOB, and Social Security Number (SSN) for each identity. A disagreement measure is computed between values in each corresponding attribute of the two identities being compared. An overall disagreement value between the two identities is calculated as an equally weighted sum of disagreement measures on the four attributes. If an overall disagreement value between two identity records is less than a threshold value, the algorithm suggests that the two identities match. The threshold value is determined by a training process. Experiments showed that this algorithm achieved high matching accuracy (94%).

Although they can achieve good performance, weighted-sum decision models have three major problems. First, a similarity score between two identity records is difficult to interpret. People intend to think of similarity scores as belief values. However, those scores sometimes do not reflect human belief. For example, the similarity score between two identical names should be one (i.e., 100% similar). If the name is a common one (e.g., John Smith), the belief that the two names refer to the

same person would be much lower than 100%. Second, weighted-sum decision models lack of a mechanism to handle missing values, which is a very common problem in record management systems. Third, the weighted-sum decision model requires a labeled training dataset for supervised learning. However, manually generating a training dataset is often very time-consuming and labor-intensive.

Originated in the area of statistics, record linkage (RL) is a probability-based decision model that determines if a pair of records describes the same entity. A formal definition for RL was given by Fellegi and Sunter (Fellegi and Sunter, 1969). In RL, a comparison between two records is represented by a feature vector γ that contains comparison features. Each comparison feature is a similarity measure of the two values on an individual attribute. Record linkage calculates an odds ratio *R* for each feature vector: $R = m(\gamma)/u(\gamma)$, where $m(\gamma)$ is the probability that the γ belongs to the matched set and $u(\gamma)$ is the probability that γ belongs to the non-matched set. The two probabilities are estimated by supervised learning. If two records match, their matching probability, $m(\gamma)$, should be greater than $u(\gamma)$, which is the probability that they do not match. Therefore, the odds ratio *R* of two matched records should be greater than that of two non-matched records. Two threshold ratios need to be determined by a supervised learning. If the ratio *R* of a comparison feature vector is greater than the upper threshold, the comparison is classified into the matched set. If the ratio *R* is smaller than the lower threshold, the comparison is classified into the non-matched set. The advantage of RL is the use of probabilities that human can easily translate into belief. However, it still needs to manually generate training datasets for supervised learning.

Ravikumar and Cohen (2004) proposed a 3-layer hierarchical graphical model for record linkage problems (Ravikumar and Cohen, 2004). In this model, a layer of latent variables were added between the class variable and the feature vector. As shown in Figure 1, X_i is a binary latent variable that represents an intermediate match decision on the value-pair of feature F_i. This model captures the intuition that a record match decision is often dependent on match decisions of features rather than on feature value comparisons directly. A multinomial probability Bayesian network (BN) is used to capture probabilistic dependencies among graph nodes. Estimating the BN parameters is accomplished by the EM algorithm. The EM algorithm can be used for fully unsupervised learning or semi-supervised learning, both of which are preferable to supervised learning. Experiments showed that this graphical model based approach with unsupervised learning achieved performance comparable to that of fully supervised record-linkage methods. However, semi-supervised learning was preferred to fully unsupervised learning because unlabeled data alone are insufficient (Nigam, McCallum, Thrun and Mitchell, 2000) and noise is subject to overfitting (Ravikumar *et al.*, 2004). In semi-supervised learning, only a part of the training data needs to be labeled.

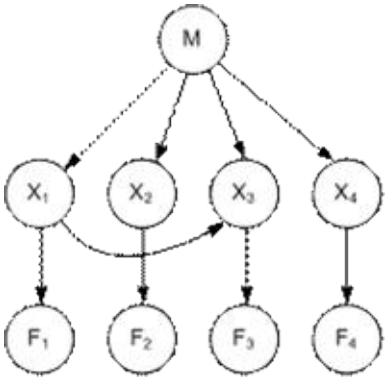

Figure 1. A 3-layer Hierarchical Graphical model

Similar to record linkage, the 3-layer hierarchical graphical model is also based on probability estimates, which is easy to interpret. It has a built-in mechanism for handling missing values. If a feature value is missing in a comparison, the BN inference algorithm will replace the missing value with an "estimated value" by considering a set of possible values associated with corresponding probability estimates. Semi-supervised learning is preferable to supervised learning because it reduces the workload in manually generating labeled training data.

In summary probability-based decision models are preferable to weighted-sum models for identity matching because of its easy translation into human belief, the built-in mechanism for handling missing values, and the semi-supervised learning.

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. 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, ID numbers (SSN and Driver's License,), gender, race, weight, height, address, and phone number. 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).

Data Collection

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 (Wang *et al*. 2004) 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 manually verified the 10 possible matches for each of the 200 suspects. 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 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 missing values in many attributes and we were unable to make a call, we would categorize it as U (uncertain).

Category	Description
D	Intentional Deception
E	Unintentional Errors
N	Not a match
U	Uncertain (too little information to make a call)

Table 1: Categories into which possible matches are classified

Taxonomy of Identity Problems

To our surprise, more than half (55.5%) of the 200 suspects had either a deceptive or an erroneous counterpart existing in the sample data. 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. The detailed statistics are shown in Table 2.

Categories	Number of Suspects	Percentage
Having a true match	111	55.5%
-- Intentional deception (D)	59	29.5%
-- Unintentional errors (E)	84	42.0%
Having no true matches (N)	56	28.0%
Having an unknown record	64	32.0%
Total	**200**	**100.0%**

Table 2: Statistics on matching identities

As shown in Figure 2, we created 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. Erroneous identities were

mostly 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. 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.

Most of the changes made in erroneous identities were minor. For example, erroneous DOB and ID numbers had only a 1-digit difference with the corresponding true values in most cases. Deception was found in values for name, DOB, ID numbers and address. Attribute values were altered more drastically in deception than in errors. We found that people preferred telling a half-truth lie to completely making up things. In most of the cases, altered values looked very similar to the corresponding true values. Although in some cases an altered value could be very different in one attribute (for example, using someone else's name), other attribute values such as DOB and ID numbers may still remain similar and help to recognize the deceptive identity.

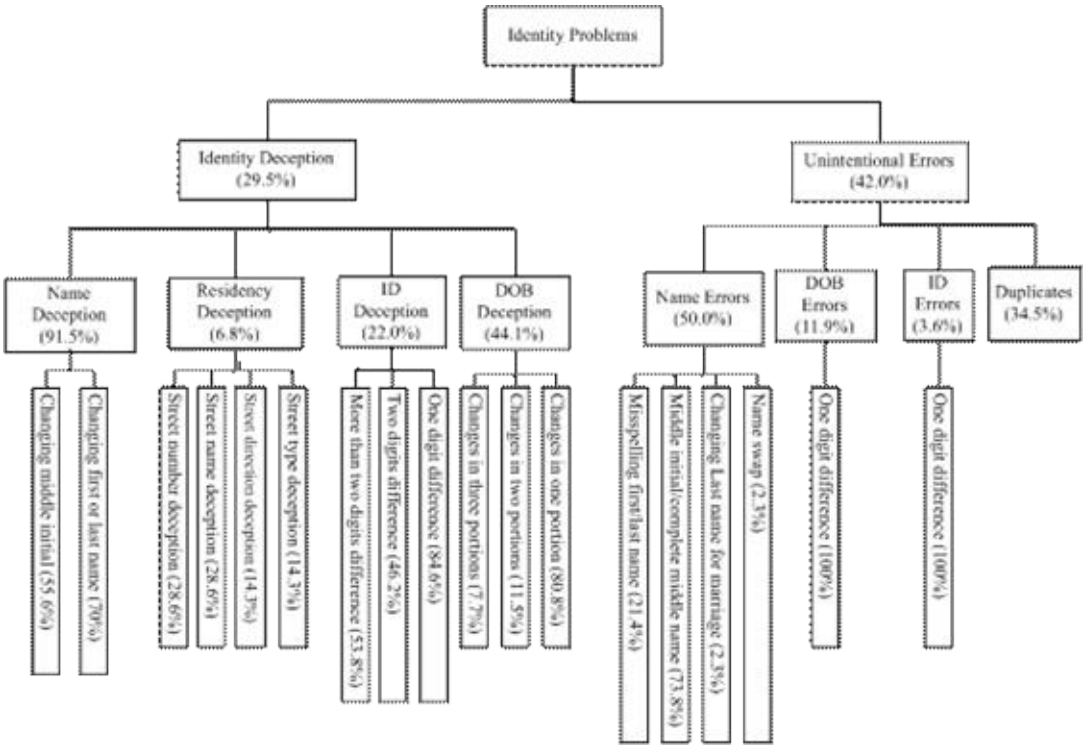

Figure 2. Taxonomy of Identity Problems

RESEARCH DESIGN

A Multi-Layer Graphical Model for Identity Matching

The taxonomy of identity problems shows that attribute values in both erroneous and deceptive identities exhibit similarity to their corresponding true values. Techniques that improve identity matching need to be able to locate identity information in an approximate rather than exact manner. Both weight-sum and probability-based decision models can be used for approximate matching. Because of its advantages in handling missing values, semi-supervised learning and easy interpretation, a probability-based decision model is preferable.

We extend the model proposed by (Ravikumar *et al.*, 2004) to a more generic one (Figure 3) that allows more than three layers, depending on the complexity when determining the agreement of values on each attribute. An identity record *I* is a

vector of attribute values, denoted as $I = \{t_1, t_2, ..., t_m\}$. Given two identity records I_a and I_b, a comparison vector is defined as a vector of comparison features: $f(I_a, I_b) = \{f_1(t_{1a}, t_{1b}), f_2(t_{2a}, t_{2b}), ..., f_m(t_{ma}, t_{mb})\}$. Each comparison feature is a similarity measure between an attribute value-pair. A binary-valued node x_i is defined as a latent match variable for attribute i $(1 \le i \le m)$. x_i equals to one if the value pair for attribute i matches or zero otherwise. The match-class variable M depends on those intermediate latent variables.

In some cases, a latent match variable x_i may have its own latent variables. For example, a match decision on a pair of names may depend on match decisions on first names, middle names, and last names. To accommodate those cases in our graphical model, sub latent match variable x_{i1}, x_{i2}, …, x_{in} can be defined for variable x_i. Each sub latent match variable depends on a similarity measure on a sub attribute value-pair (e.g., a pair of first names).

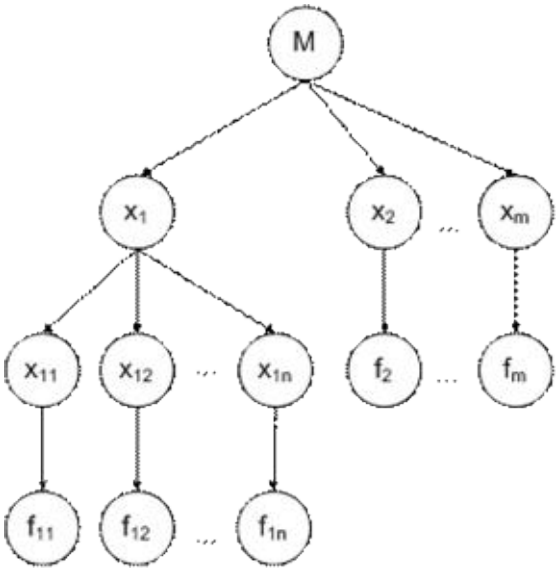

Figure 3. A Multi-Layer Graphical Model

As shown in the taxonomy of identity problems, values of name, DOB, address and ID numbers often indicate a match between two identities. Therefore, we include those four attributes in our proposed graphical model for identity matching (Figure 4). A match decision on a pair of names depends on match decisions on its sub attributes such as first name, middle name, and last name. Match decisions on all other attributes depend directly on the similarity measures of the corresponding value-pairs. A similarity measure of a value-pair is calculated using Edit Distance (Levenshtein, 1966), which counts the minimum number of insertion, deletion, and substitution required to transform one string to the other. Given two strings S_1 and S_2, the similarity measure is defined as:

$$Sim(S_1, S_2) = 1 - \frac{ED(S_1, S_2)}{max(|S_1|, |S_2|)}$$

where *ED()* is the Edit Distance function and $|S|$ calculates the length of string S.

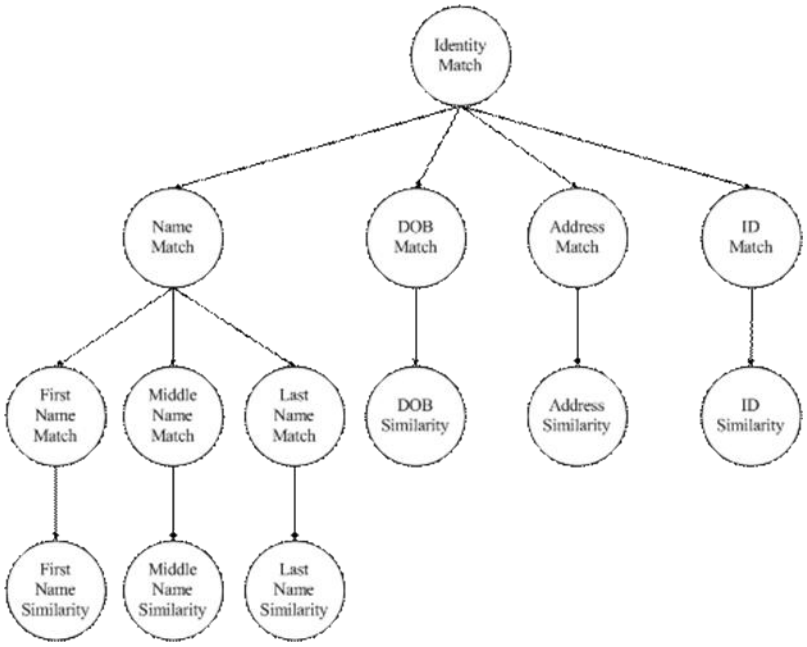

Figure 4. A Graphical Model for Identity Matching

Bayesian Network Learning and Inference

We use Bayesian network (BN) to estimates parameters such as prior probabilities (e.g., P(*Sim*(DOB)=1) and conditional probabilities (e.g., P(*Sim*(DOB)=1|DOB match=1)) for the proposed graphical model. These parameters will be later on used in Bayesian network inference algorithms. A BN inference algorithm infers an identity match decision given a vector of similarity measures and a parameterized BN. We use Maximum Likelihood Estimation (MLE) (Spiegelhalter and Lauritzen, 1990) for BN learning and Pearl's polytree propagation algorithm (Pearl, 1988) for BN inference.

Bayesian network is a multinomial probability model. Continuous feature values such as similarity scores cannot be directly used in BN learning and inference algorithms. One technique, called discretization, discretizes a continuous feature value domain into a number of intervals (Ravikumar *et al.*, 2004). Feature values that fall in the same interval have the same nominal value. Continuous feature values are converted to nominal feature values that can be used in BN learning and inference algorithms. However, if the number of intervals is too small, discretization will lead to a poor approximation of the continuous distribution. If the number is too large, it will increase the computational complexity of both BN learning and inference algorithms.

Semi-supervised is preferable to fully supervised learning because manually labeling training data is often time-consuming and labor-intensive. We use the Expectation-Maximization (EM) algorithm (Lauritzen, 1995) to conduct BN semi-supervised learning. The first step of the EM algorithm is the prime *M* step. In semi-supervised learning, the prime *M* step uses MLE to estimate the BN parameters θ with a small set of labeled training data. Given the parameters estimated in the prime *M* step, step *E* computes the expected values of unknown class labels in the training dataset using a BN inference algorithm. Step *M* treats the expected values as though they were observed and estimate a new set of BN parameters θ using MLE. Steps *E* and *M* iterates until the likelihood score converges. The likelihood function is defined as:

$$p(D \mid \theta) = \prod_{i=1}^{N} p(x_i \mid x_{parent(i)})$$

where *D* is the training dataset, *N* is the number of records in *D*, x_i is a nominal feature value, $x_{parent(i)}$ is a nominal feature value of x's parent node.

The BN learning algorithm may suffer a sparse data problem when training data are not sufficient. This problem is caused by discretization. The training data may not contain values in certain feature value intervals. For example, the learning algorithm may get the following two parameters: $p(Sim(FN) = 0.6 \mid FNMatch = 1) = 0.5$ and $p(Sim(FN) = 0.7 \mid FNMatch = 1) = 0$ from learning. However, the latter probability is expected to be greater than the former one. Ravikumar and Cohen (Ravikumar *et al.*, 2004) enforced a monotonic increasing constraint in parameter estimation to solve this problem. We adopt their idea and use a heuristic-based approach to simplify the computation. Given two nominal feature values x_i and x_j, we define the monotonic increasing as the following:

$$\begin{cases} P(x_i \mid Match = 1) \geq P(x_j \mid Match = 1), \text{if } x_i \geq x_j \\ P(x_i \mid Match = 0) \leq P(x_j \mid Match = 0), \text{if } x_i \geq x_j \end{cases}$$

With the monotonic increasing constraint, the latter probability in the example becomes $p(Sim(FN) = 0.7 \mid FNMatch = 1) = 0.5$.

EXPERIMENTS

We conducted experiments using real law enforcement data to examine the performance of our proposed graphical model for identity matching. An exact-match based technique is our baseline method.

Dataset

The data collected for the case study were used again in our experiments. There are 200 primary identities in the dataset. For each primary identity there are 10 possible matches included in the dataset. We only use 5 attributes that are necessary for the proposed graphical model: name, DOB, address, SSN, and Driver's License Number. With the help of a police detective veteran, we manually examined each of the possible matches and categorized it into one of the four categories defined in Table 1.

We generated the training dataset by comparing each primary identity to every one of its possible matches. In total we had 2,000 (200*10) identity comparisons. Each comparison is represented by a feature vector and a class label. The feature vector consists of similarity scores between corresponding attribute values. The ID similarity score took the greater value between the similarity score of SSNs and that of Driver's License Numbers. The class label was assigned one if the possible match was in the category of deception or error, and zero otherwise.

Performance Matrices

We consider identity matching as a classification problem. When we compare a class label predicted by a matching algorithm to the actual class label, the result falls in one of the four categories defined in Table 3.

	Actual Class Label=1	**Actual Class Label=0**
Predicted Class Label =1	True Positive (TP)	False Positive (FP)
Predicted Class Label =0	False Negative (FN)	True Negative (TN)

Table 3. Categories of Classification Results

We evaluate the algorithm's classification accuracy by using three kinds of measures: recall, precision, and F-measure. Those measures are widely used in information retrieval (Salton, 1988). Precision, in this scenario, is defined as the percentage of correctly detected matched identities in all matched identities suggested by the algorithm. Recall is the percentage of matched identities that are correctly identified. F-measure is a well-accepted single measure that combines recall and precision.

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$f-measure = \frac{2 * precision * recall}{precision + recall}$$

Experimental Results

In our experiment we compared the performance of the proposed graphical model for approximate identity matching to that of the exact-match technique often used in information systems.

To evaluate the performance of the exact-match technique, we compared each identity to every other identity in the training dataset. The predicted class label for each comparison was determined by a heuristic approach commonly used in law enforcement communities (Marshall, Kaza, Xu, Atabakhsh, Petersen, Violette and Chen, 2004). The class label of a comparison was assigned one (i.e., a match) only if the first name, the last name and the DOB value of one identity were identical to those of the other identity being compared.

We used a 10-fold cross-validation method in evaluating the performance of the proposed graphical model. We first randomly divided the training dataset into 10 folds. Each time we used 9 folds for training and used the other fold for testing.

We also evaluated the performance of the proposed graphical model using different learning methods such as supervised and semi-supervised learning. In semi-supervised learning we manipulated the ratio of unlabeled data in the training dataset by randomly removing the class labels.

Table 4 summarizes the experimental results. The proposed graphical model performed significantly better than the exact-match technique originated from human heuristic. When 90% of training data were unlabeled, the graphical model was not able to find matching identities. Precisions were high (≥0.963) and steady when the ratio of unlabeled training data *R* was less than 0.9. Recalls ranged from 0.728 to 0.970. When the ratio R is set to 0.3, the proposed graphical model performed the best (F-measure=0.984). The semi-supervised learning still performed better than the fully supervised learning when only 20% of training data were labeled. The performance of the graphical model with semi-supervised learning is also shown in Figure 5.

			Precision	**Recall**	**F-Measure**
Exact-Match			0.421	0.381	0.400
Proposed Graphical Model	Supervised Learning		0.996	0.728	0.838
	Semi-Supervised Learning	R=0.1	0.986	0.740	0.843
		R=0.2	0.982	0.907	0.942
		R=0.3	**1.000**	**0.970**	**0.984**
		R=0.4	1.000	0.908	0.951
		R=0.5	0.992	0.921	0.954
		R=0.6	1.000	0.964	0.980
		R=0.7	1.000	0.850	0.899
		R=0.8	0.963	0.744	0.789
		R=0.9	0.000	0.000	0.000

Table 4. Experimental Results (R is the ratio of unlabeled data in training data)

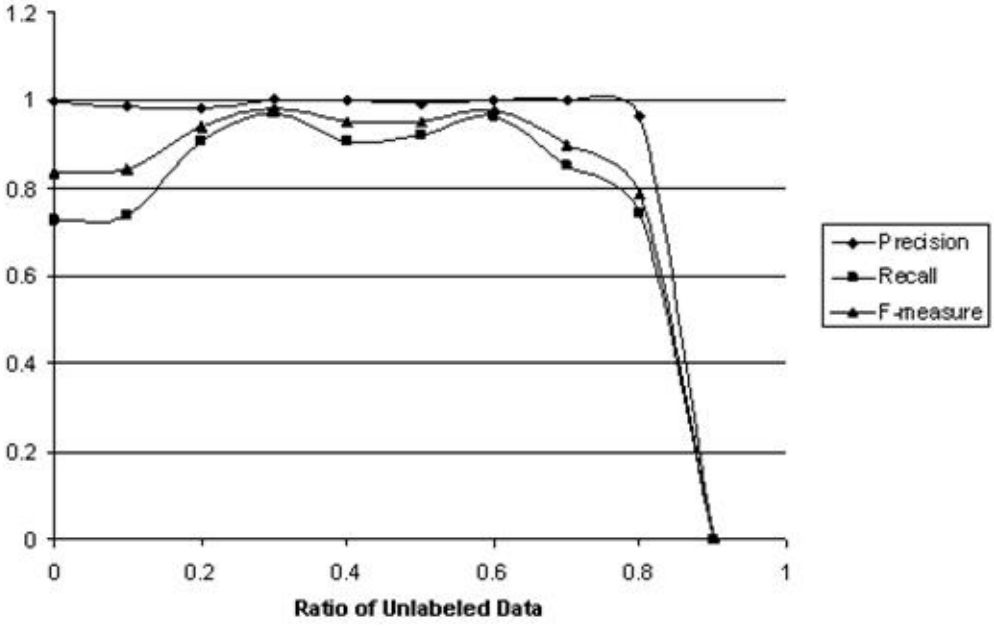

Figure 5. Performance of the Graphical Model with Semi-Supervised Learning

CONCLUSION

In this study we examine the identity problems that result in difficulties for matching identity information. Two types of problems, including intentional deception and unintentional errors, were found in real law enforcement records. We created a taxonomy of identity problems based on findings.

We proposed a multi-layer graphical model for approximate identity matching. Preliminary experiments showed that it performed significantly better than the exact-match technique. The proposed model does not require a fully labeled training dataset. With only 20% of labeled data, the model achieved performance comparable to that of the supervised learning. We will compare our proposed model with more sophisticated techniques such as record linkage in future research.

REFERENCES

1. Ba, S., Lang, K. R., and Whinston, A. B. (1997) Enterprise Decision Support Suing Intranet Technology, *Decision Support Systems* 20), 99-134.
2. Badiru, A. B., Karasz, J. M., and Holloway, B. T. (1988) AREST: Armed Robbery Eidetic Suspect Typing Expert System, *Journal of Police Science and Administration* 16), 210-216.
3. Bolloju, N., Khalifa, M., and Turban, E. (2002) Integrating Knowledge Management into Enterprise Environments for the Next Generation Decision Support, *Decision Support Systems* 33), 163-176.
4. Brown, D. E., and Hagen, S. (2002) Data Association Methods with Applications to Law Enforcement, *Decision Support Systems* 34, 4), 369-378.
5. Clarke, R. (1994) Human Identification in Information Systems: Management Challenges and Public Policy Issues, *Information Technology & People* 7, 4), 6-37.
6. Dey, D., Sarkar, S., and De, P. (2002) A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases, *IEEE Transactions on Knowledge and Data Engineering* 14, 3), 567-582.
7. Fellegi, I. P., and Sunter, A. B. (1969) A Theory for Record Linkage, *Journal of the American Statistical Association* 64, 328), 1183-1210.

8. GAO "Law Enforcement: Information on Timeliness of Criminal Fingerprint Submissions to the FBI," GAO-04-260, United States General Accounting Office (GAO).
9. HomeOffice, U. K. "Identity Fraud: A Study," United Kingdom HomeOffice.
10. Kent, S. T., and Millett, L. I. (2002) IDs--Not that Easy: Questions About Nationwide Identity Systems National Academy Press, Washington, D.C.
11. Lauritzen, S. L. (1995) The EM Algorithm for Graphical Association Models with Missing Data, *Computational Statistics & Data Analysis* 19, 1995), 191-201.
12. Levenshtein, V. L. (1966) Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, *Soviet Physics Doklady* 10), 707-710.
13. Marshall, B., Kaza, S., Xu, J., Atabakhsh, H., Petersen, T., Violette, C., and Chen, H. (Year) Cross-Jurisdictional criminal activity networks to support border and transportation security, in (Ed.)^(Eds.), *The 7th Annual IEEE Conference on Intelligent Transportation Systems (ITSC 2004)*, Washington, D.C.
14. Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000) Text Classification from Labeled and Unlabeled Documents using EM, *Machine Learning* 39), 103-134.
15. Pearl, J. (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference Morgan Kaufmann Publishers, San Mateo, CA.
16. Ravikumar, P., and Cohen, W. W. (2004) A Hierarchical Graphical Model for Record Linkage, in: *Conference on Uncertainty in Artificial Intelligence (UAI '04)*, Banff Park Lodge, Banff, Canada.
17. Salton, G. (1988) Automatic text processing: the transformation, analysis, and retrieval of information by computer Addison-Wesley Pub.
18. Spiegelhalter, D., and Lauritzen, S. (1990) Sequential Updating of Conditional Probabilities on Directed Graphical Structures, *Networks* 20), 579-605.
19. Wang, G., Chen, H., and Atabakhsh, H. (2004) Automatically Detecting Deceptive Criminal Identities, *Communications of the ACM* 47, 3), 71-76.