Journal of the Association for Information Systems

JAIS

Special Issue

Complex Problem Solving: Identity Matching Based on Social Contextual Information *

Jennifer Xu
Computer Information Systems
Bentley College
jxu@bentley.edu

Michael Chau
School of Business
The University of Hong Kong
mchau@business.hku.hk

G. Alan Wang
Business Information Technology
Virginia Polytechnic Institute and State University
alanwang@vt.edu

Jiexun Li
College of Information Science and Technology
Drexel University
jiexun.li@ischool.drexel.edu

Abstract:

Complex problems like drug crimes often involve a large number of variables interacting with each other. A complex problem may be solved by breaking it into parts (i.e., sub-problems), which can be tackled more easily. The identity matching problem, for example, is a part of the problem of drug and other types of crimes. It is often encountered during crime investigations when a single criminal is represented by multiple identity records in law enforcement databases. Because of the discrepancies among these records, a single criminal may appear to be different people. Following Enid Mumford's three-stage problem solving framework, we design a new method to address the problem of criminal identity matching for fighting drug-related crimes. Traditionally, the complexity of criminal identity matching was reduced by treating criminals as isolated individuals who maintain certain personal identities. In this research, we recognize the intrinsic complexity of the problem and treat criminals as interrelated rather than isolated individuals. In other words, we take into consideration of the social relationships between criminals during the matching process. We study not only the personal identities but also the social identities of criminals. Evaluation results were quite encouraging and showed that combining social features with personal features could improve the performance of criminal identity matching. In particular, the social features become more useful when data contain many missing values for personal attributes.

Keywords: *Complex problems, design science, identity matching, social contextual information.*

** This is a part of the special issue on Enid Mumford's contribution to information systems theory and theoretical thinking. Jaana Porra and Rudy Hirschheim were the accepting guest editors.*

Volume 8, Issue 10, Article 2, pp. 525-545, October 2007

Introduction

Modern society is increasingly facing various complex problems that are "pervasive, spreading unhindered into regions, countries, and economic activities which seem powerless to resist the invasion" (Mumford, 1998, p. 447). Globalization, for example, is such a complex problem that, while bringing numerous opportunities to organizations, has also brought substantial challenges and pressure. Defining complex problems seems to be a good starting point for solving them; however, there has not been a widely accepted definition (Gray, 2002; Quesada et al., 2005). Funke (1991) suggested that complex problems can be understood by contrasting them with simple problems, which can be solved by simple reasoning and pure logic (Quesada et al., 2005), and that they can be characterized by their intransparency, polytely (from the Greek words *poly telos* meaning many goals), complexity, connectivity of variables, dynamic, and time-delayed effects. In other words, the defining characteristics of complex problems are a large number of variables (complexity) that interact in a nonlinear fashion (connectivity), changing over time (dynamic and time-dependent), and to achieve multiple goals (polytely).

One of the two examples that Mumford used to illustrate complex problem solving (Mumford, 1998; Mumford, 1999) is drug crimes, which clearly have all these features: a large number of people interact and cooperate frequently; they play different roles and spread across different countries and regions; they form networks of personnel to carry out various activities (drug production, transportation, distribution, sales, and money laundering); they change from time to time in response to the uncertainty and dynamics in their environments; their goals are to effectively and efficiently maximize profit and minimize damage and loss. Although drug crimes are not directly related to many organizations, they are likely to become one of our society's major problems that will have social, health, and economic impact on our lives (Mumford, 1998).

Solving drug crimes is by no means an easy task. Like many other complex problems, the drug problem consists of many sub-problems, which themselves are also complex. *Identity matching* is such a sub-problem of drug crimes. This problem is often encountered during investigations of an organized crime (e.g., drug trafficking, arms smuggling, and money laundering) or serial crimes (e.g., serial fraud and serial sex offenses). Many criminals, especially drug barons and dealers, often disguise their identities by providing misleading or deceptive information (e.g., fake names and identification numbers). As a result, a single criminal may appear to be two distinct people in two different cases, making it difficult for crime investigators to link the two cases together.

Effectively matching criminal identities is important because it helps enhance the information sharing, collaboration, and coordination abilities of crime investigators, law enforcement, and security agencies at different levels. It allows these government entities to consolidate information from different sources, identify new investigative leads, and perform further analysis by connecting seemingly unrelated cases—one of the most important processes in crime analysis (Brown and Hagen, 2002; Chen et al., 2003; Ianni and Reuss-Ianni, 1990), and develop disruptive strategies to break criminal organizations. This would largely facilitate the effort to combat the problem of drugs as well as other crimes.

Information technology has played a critical role in tackling the identity matching problem. Various techniques have been proposed. Unfortunately, the effectiveness of these techniques still needs to be improved. In this paper, we design a new method for tackling the problem of matching criminal identities. Traditionally, the complexity of criminal identity matching was reduced by treating criminals as isolated individuals who maintain certain personal identities. Such an approach is too simplistic and does not consider the relationships between criminals who are involved in the illegal drug dealing and trafficking processes. As Mumford pointed out, drug crimes must be viewed from a network perspective (Mumford, 1998). In this research, we recognize the intrinsic complexity of the identity matching problem and treat criminals as interrelated rather than isolated individuals. Each individual plays one or more parts, be it a dealer, a smuggler, or a drug user, in a large social network. In other words, we take into consideration the social relationships between criminals and study not only the *personal identities* but also the *social identities* of criminals. We hope to find out whether social contextual information can help improve the effectiveness of identity matching techniques and whether these additional features become more useful when the data quality is low.

Our research can be understood in terms of Mumford's three-stage problem solving framework (Mumford, 1998): seeing the total picture, developing strategies, and taking action. The framework is general enough to provide a guideline for solving any complex problem. We position our research in the third stage, in which we attempt to operationally address the identity matching problem, a specific facet of the drug crime problem.

The rest of the paper is structured as follows. First we review related literature on the concept of identity and existing identity matching techniques. Research questions are then raised. In the next section, we present our research design and propose an identity matching method using both personal and social features. We then report on the evaluation studies for assessing the effectiveness of the proposed method based on real drug crime data. The identity matching problem is then discussed in the larger context of the illegal drug problem using Mumford's three-stage framework. Finally, we conclude our paper with some discussion on the limitations, implications, and future work of our research.

Theoretical Background and Related Research

Theories of Identity

The concept of identity has long been studied in philosophy, psychology, and sociology. Identity generally has two basic aspects: personal identity and social identity. Personal identity is defined as one's self-perception as an individual (Cheek and Briggs, 1982). It deals with the necessary and sufficient conditions under which self persists over time. For example, people often ask common questions about their personal identities: Who am I? Where did I come from?

The theories of social identity diverge between the psychological view and sociological view. The psychologically-based theory of social identity (PSIT) deals with the cognitive and psychological process of an individual's self-perception as a member of certain labeled categories (Tajfel and Turner, 1986; Turner, 1999), including a nationality, culture, ethnicity, gender, and employment. The sociologically-based identity theory (SSIT), on the other hand, "focuses on the relationships between social actors who perform mutually complementary roles (e.g., employer-employee, doctor-patient)" (Deaux and Martin, 2003, p. 102). The emphasis is on the interpersonal relationships between people and the social structure and context formed based on the relationships (Stryker and Serpe, 1982). The social context determines the specific roles an individual takes. For example, a man can take different roles in his family: the father of his children, the son of his parents, and the husband of his wife. An individual's social identity, in this sense, is defined by the role-based interactions between the individual and the surrounding people (Stryker and Serpe, 1982).

Research on the concept of identity provides a sound theoretical foundation for our study. Although these theories do not explicitly indicate which features can be used in the identity matching problem, they point to the directions in which useful information can be found to tackle the problem. Based on these theories, we categorize identity information into *personal information* and *social contextual information*.

For both types of information, our interest is in the features that can be used to practically distinguish an individual from others. Clarke (1994) listed a number of personal features that can be used in human identification such as name, physical characteristics, and appearance. These personal features can be categorized into four types: *given identity features*, *physical characteristics*, *biometric features*, and *biographical features*. Given identity features are identifiers assigned to an individual at birth, such as name, date of birth (DOB), place of birth, mother's maiden name, and Social Security Number (SSN). Physical characteristics include weight, height, hair color, eye color, and visible physical marks such as tattoos. Biometric features include characteristics that are unique to an individual such as fingerprints, DNA, iris, hand geometry, and voice, among many others. Information that builds up over an individual's lifespan comprises the individual's biographical identity, examples of which are education and employment background, credit history, medical history, crime history, etc.

For social identity, SSIT is more relevant to our research because our interest is not in the psychological process of self-perception but in the external features of social identity. We focus on the proximate social groups of individuals. A social group around an individual is defined by people directly interacting with him/her. The social contextual information, which includes the social structure of the group, the relationships between the individual and other members, and the roles the individual takes, is used for defining the individual's social identity.

In reality, different kinds of personal information vary in availability and reliability. For example, identity records stored in law enforcement databases often only contain individuals' simple given identity features, physical characteristics (e.g., weight and height), and sometimes biometric features. The given identity information is subject to deception and many other data quality issues. The physical characteristics are not reliable since they often can be easily altered. Hair color, for example, can be changed from time to time. Although biometric features such as fingerprints and DNA are the most difficult to falsify and can reliably identify an individual, they are rarely available.

The social identity of an individual, in contrast, usually cannot be easily altered or falsified because such information is embedded in the social context formed through the interactions of group members. The social contextual information is expected to provide additional information for distinguishing an individual from others. Thus, our first research question is:

RQ1: Can we use social contextual information to help match criminal identities?

Existing Identity Matching Techniques

The effectiveness of existing identity matching techniques is far from satisfactory. One important reason is that they utilize only personal features that are available in current record management systems used by most law enforcement and intelligence agencies. The success of these techniques, to a large extent, relies on the high quality of data. When data quality is low due to deception (Wang et al., 2004), errors (Redman, 1998), or missing values, personal features cannot provide sufficient and correct information for matching identities accurately and effectively.

Based on the way that a matching decision model is constructed, existing techniques can be categorized into two types: heuristic techniques and machine learning techniques. Heuristic techniques often rely on domain experts to manually specify decision rules. In a study on cross-jurisdictional information integration, Marshall et al. (2004) provided a simple identity matching heuristic based on domain experts' suggestions. The heuristic considers two identity records as a match only if their first name, last name, and DOB values are identical. This method is subject to a high rate of false negatives. Due to data quality issues, it is very likely that identity records referring to the same individual may have disagreeing values in any of the three attributes. The IBM DB2 Identity Resolution (EAS), an advanced heuristic matching technique, is a leading commercial product designed to manage identity records (Jonas, 2006). A resolution or matching score is calculated for a pair of identity records using a set of rules pre-defined by domain experts. Matching decisions are made based on the score with certain rules. For example, if the DOB and last name values of two identity records are identical and the matching score of their first names is above 70, the two records are resolved into one. Another example of the decision rules is that if the overall resolution score of two identity records is greater than 100, they are resolved to the same individual. Such a rule-based technique relies heavily on experts' involvement in defining the rules for satisfactory matching performance. The rule-defining process often is very time-consuming, and the rules have low portability in different settings. Thus, the applicability of heuristic approaches is rather limited in practice.

A machine learning technique automatically builds a decision model by learning the parameters in the model from a training dataset. The training set consists of pairs of records that have already been classified as match or non-match. Compared with heuristic techniques, they are quite efficient, with less or no human intervention. Machine learning techniques compare two records by individual features, and the decision model requires the calculation of a similarity score between the records based on these features. Dey et al. (2002), for example, proposed an integer programming approach for entity reconciliation. The objective function of the model is to minimize the total cost of type-I and type-II errors in matching decisions based on similarity scores. The problem with this approach is that it assumes that one entity from a data source can be matched to one and only one entity in the other data source. However, this assumption is rarely true in the real world. Brown and Hagen (2002) proposed a data association method for linking criminal records that possibly refer to the same suspect. This method compares two records and calculates a total similarity score as a weighted sum of the similarity scores of all corresponding feature values of the two records. This method makes use of various features such as hair color, eye color, and other physical characteristics. However, it does not provide a decision model based on which a matching decision can be made.

Wang et al. (2004) proposed a record comparison algorithm for detecting deceptive criminal identities. It uses four personal features: name, DOB, SSN, and address. A normalized Euclidean function is used to calculate the overall similarity score. Two records are considered a match if the similarity score is higher than a pre-defined threshold. Experiments showed that this technique was effective in matching identity records.

Although machine learning techniques are more efficient than heuristic approaches, learned decision models may be flawed due to factors such as low data quality. Wang et al. (2006) revealed that missing data could significantly affect the performance of the record comparison algorithm. Incomplete records with many missing values can be mistakenly matched up, resulting in a higher error rate. For example, two records that both have only "John" recorded as the first name and values of all other features (e.g., last name, DOB) missing would be considered a match by the decision model. This is a common limitation of many identity matching techniques utilizing only personal features.

In summary, identity matching is a complex problem as it deals with many different limitations such as criminal deception and data quality issues. More importantly, criminals are not isolated individuals but relate to and interact with one another. Existing identity matching techniques that use only personal features cannot effectively tackle the identity matching problem when the data quality is poor. Our second research question deals with the effectiveness of our method when data quality varies:

> RQ2: *Under what circumstances* (*e.g., levels of data quality*) *does social contextual information show more effectiveness for identity matching*?

Research Design

We follow the design science methodology presented in Hevner et al. (2004) because design science is inherently a problem solving paradigm. Unlike natural science that aims at developing laws and theories that "make claims about the nature of reality," design science "attempts to create things that serve human purposes." The key question to ask in design science is not "how and why things are" but "does it work?" or "is it an improvement?" (March and Smith, 1995, p. 253). The goal of design science research is to address "unsolved problems in unique or innovative ways or solved problems in more effective and efficient ways" (Hevner et al., 2004, p. 81). This goal is achieved by building and applying the designed artifacts, which may be constructs, models, methods, and instantiations (March and Smith, 1995). Our research is aimed at addressing the identity matching problem that has not been completely solved. The artifact we have built is primarily a method that improves existing techniques innovatively. Both our research questions are intended to answer one fundamental question: does the method work?

Hevner et al. (2004) provided seven guidelines for design science research: (1) design as an artifact, (2) problem relevance, (3) design evaluation, (4) research contributions, (5) research rigor, (6) design as a search process, and (7) communication of research. Fundamentally, design is both a product (artifact) and a process (build and evaluate). In other words, what is essential to design science research is to design an artifact (Guideline 1) through the construction (Guideline 6) and evaluation (Guideline 3) process. We have already discussed the relevance of the identity matching problem (Guideline 2) in the context of drug crimes. In this and the next section we will focus on the process and product of our design following Guidelines 1, 3, and 6. Research rigor (Guideline 5), contributions (Guideline 4), and communication (Guideline 7) will be discussed in the last section.

Our design artifact (method) can be considered to be a machine learning technique. It consists of four major components: *feature extraction*, *person record clustering*, *within-cluster pair-wise comparison*, and *classification*. Figure 1 illustrates how these four components work together to match criminal identity records. These four components can also be viewed as three processes: search (or extract), compare, and decide. In the first process, important features (personal and social) are identified and extracted from criminal records. In the comparison process, which consists of the clustering and pair-wise comparison components, the extracted features are used to find out how similar two originally unrelated records are. In the last process, the decision to classify the two records as either a match or a non-match is made based on a similarity score resulting from the comparison process.

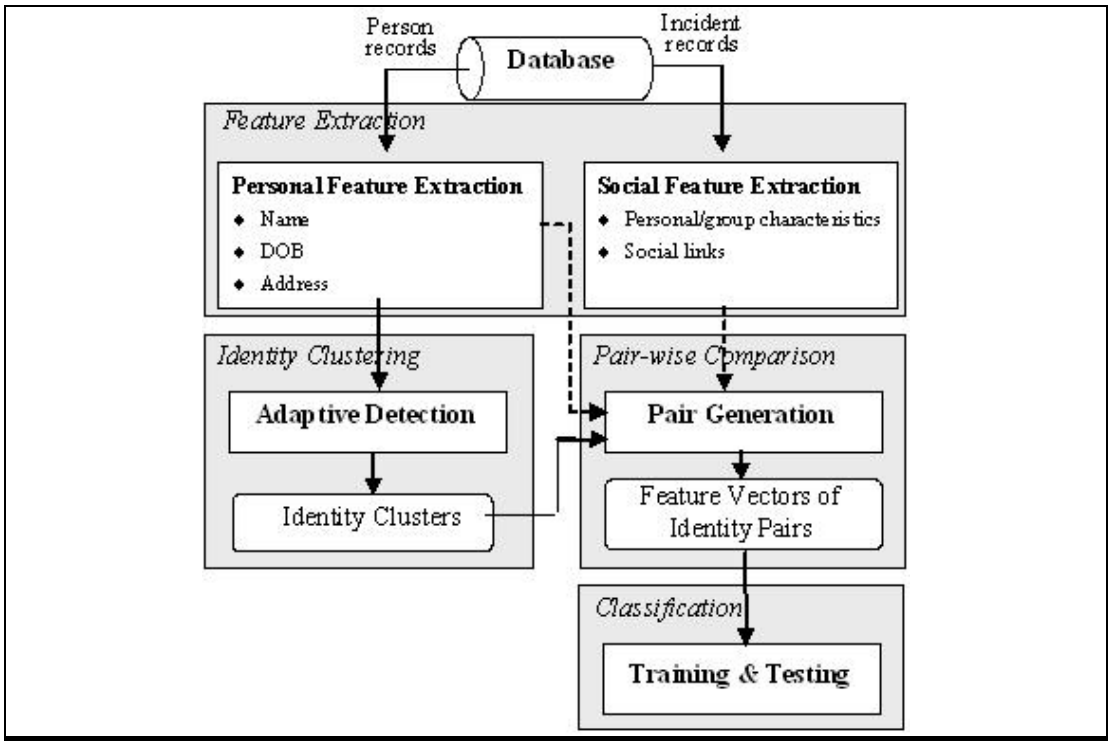

Figure 1. The Four Components in the Proposed Approach

Journal of the Association for Information Systems

Feature Extraction

Feature extraction is basically a search process in which we reduce the solution space by identifying and constructing features, given available data, which can most effectively match those seemingly unrelated criminal records. We leverage information from two basic types of records in law enforcement databases: person records and crime incident records. Person records usually contain basic given identity information (e.g., name, DOB, gender), physical characteristics (e.g., height, weight), and sometimes biometric data such as fingerprints. Because biometric information is rarely available and physical characteristics often are unreliable we use only given identity features to represent personal identity information. Incident records contain information about specific crime incidents (e.g., time, place, and crime type), as well as persons involved and their roles. The role of a person in a crime can be suspect, arrestee, victim, witness, etc. In this paper, we use only three roles that are generally available in most law enforcement databases: suspect, arrestee, and victim. Since it was found in a previous study (Schroeder et al., 2007) that the suspect and arrestee types are very similar, we combine them and use suspect/arrestee to represent both.

We use two types of features in our method: personal features and social features. Personal features include first name, last name, DOB, and address in person records. Social features, which usually are not directly available in police records, are extracted from crime incident records. We consider the roles a person took, the types of crimes the person was involved in, and the relations between the person and those he/she committed crimes with. Specifically, these social features include *personal role*, *person-associated crime type*, *social link, group role*, and *group-associated crime type*.

Personal role. Although this feature contains "personal" in its name, it is actually a type of social feature. The theory of social identity (SSIT) suggests that the role an individual takes in a social context is an important indicator of his/her social identity (Deaux and Martin, 2003). The value of the personal role is derived from incident records based on the roles a person took in past crimes. Specifically, the personal role feature of a person, $f_r(i)$, is represented as a point in a two-dimensional space indicating how frequently a person played one of the two roles (suspect/arrestee or victim):

$$f_r(i) = (sa_i, v_i),$$

where sa_i (or v_i) is the number of times person i acted as a suspect/arrestee (or victim) divided by the total number of past crimes that i was involved in. Obviously, $sa_i + v_i = 1$. Because the personal role feature somewhat summarizes the crime history of a person, it can also be viewed as a type of biographical information.

Person-associated crime type. This feature is based on the types of crimes that a person was involved in. For example, one person might have been involved more frequently in drug-related crimes, while another person has been involved in more automobile thefts. This feature is defined as the percentage of different categories of crimes a person was involved in:

$$f_c(i) = (c_{i1}, c_{i2}, c_{i3}, \ldots),$$

where c_{i1}, c_{i2}, c_{i3} are the frequencies of the types of crimes person i was involved in. Again, $c_{i1} + c_{i2} + c_{i3} + \ldots = 1$. In this research we use 61 categories of crime types including narcotic drug offenses, homicide, sexual assault, robbery, etc.

Social link. Drug crimes are carried out by networked criminals. To find out who is related to a person in question, we need the social link information. Because no social link information is directly available in law enforcement databases, we use the concept space approach (Chen and Lynch, 1992) to extract co-occurrence links from crime incident records. The concept space approach is widely used in information retrieval applications (Chen et al., 1998; Hauck et al., 2001). It generates a thesaurus from documents by calculating the frequency with which two words or phrases appear in the same documents. The more frequently two words or phrases appear together, the more likely it will be that they are related terms. We treat each incident record as a document and each person's name as a phrase. We then calculate co-occurrence weights based on the frequency with which two people appear together in the same crime incident. We assume that criminals who committed crimes together might be related and that the more often they appeared together the more likely it would be that they were related. As a result, a non-zero value of a co-occurrence weight implies a link between two persons (Hauck et al., 2002).

Group role. This feature captures the characteristics of the social group surrounding a person. We consider the collection of people who are directly related to the person to be his/her social group. In other words, the characteristics of the social group of a person are represented by the characteristics of his/her immediate "neighbors." Figure 2 illustrates the social group realization. Because persons 2, 3, 5, and 6 are directly connected to person 1, they represent the social group that person 1 belongs to. Person 4, on the other hand, is not a member of this social group because he/she is not directly related to person 1.

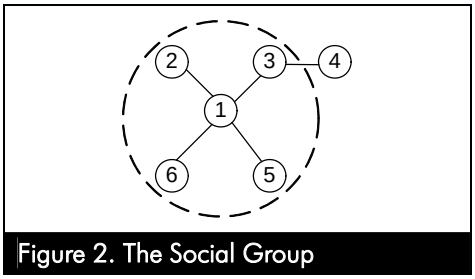

Figure 2. The Social Group

Assuming the values of personal role features of members in the social group *g* surrounding person *i* are known, the group characteristic is represented as a point in a two-dimensional space:

$$f_r^{\,g}(i) = \left(\frac{\sum_{j \in g} sa_j}{n_g}, \frac{\sum_{j \in g} v_j}{n_g} \right),$$

which is the average of group members' personal role feature values. n_g is the size of the social group.

Group-associated crime type. Similar to the group role feature, this feature is derived from the crime type profiles of the members in the group around person *i*:

$$f_c^{\,g}(i) = \left(\frac{\sum_{j \in g} c_{j1}}{n_g}, \frac{\sum_{j \in g} c_{j2}}{n_g}, \frac{\sum_{j \in g} c_{j3}}{n_g}, \ldots \right).$$

Person Record Clustering

The comparison process is used to find potentially matching records. To find matching identity records in a database, one approach is to compare every pair of records in the database. However, law enforcement databases often maintain millions of records. It is almost impossible to examine every pair of records in the database. Thus, in the clustering component, we first use an adaptive detection algorithm (Wang et al., 2006) to reduce the comparison space by filtering out obvious non-matching identity pairs based on personal features (e.g., name, DOB). The algorithm generates a list of clusters. Identity records belonging to the same cluster are considered to be candidates for matching identities. In the next component, each pair of candidates in the same cluster will be compared to find their similarity.

Because the main goal of this clustering component is to reduce the comparison space, we do not need to use all the features (including social features). In addition, the extraction of social features, especially the social links, is often very time-consuming for large databases. By reducing the comparison space, only the social links for candidate identities have to be extracted, thus significantly lowering the demand for computational resources.

In this clustering component, each person record is represented as a feature value vector, $F(i) = \{f_1(i), f_2(i), \ldots, f_k(i)\}$, where *k* is the feature index and *i* is the record index. The adaptive detection algorithm first sorts the list of identity records on a key feature such as name. The algorithm assumes: (1) matching identity records have similar values in the key feature; (2) matching identity records are located close to each other after being sorted. The algorithm examines every record in the sorted record list and compares it to its neighboring records. A window size determines the number of neighboring records that a record is compared with in the sorted list. The window size is adaptive in the sense that it increases when many matching records exist in the neighborhood. For each comparison, the algorithm calculates similarity scores for corresponding feature values. For numerical feature values, the similarity score from a feature, f_m, between records *i* and *j* is computed as:

$$Sim_m(i, j) = 1 - \frac{|f_m(i) - f_m(j)|}{\max(f_m) - \min(f_m)}.$$

The similarity between nominal features is calculated as:

$$Sim_m(i, j) = \begin{cases} 1, \text{ when } f_m(i) \text{ and } f_m(j) \text{ agree} \\ 0, \text{ otherwise} \end{cases}$$

Similarity scores of textual feature values (e.g., names) can be calculated using a string matching technique such as the Levenshtein Edit Distance (Levenshtein, 1966).

The algorithm then combines individual feature similarity scores into an overall similarity score using a normalized Euclidean function:

$$Sim(i, j) = \sqrt{\frac{Sim_1(i, j)^2 + Sim_2(i, j))^2 + ... + Sim_k(i, j)^2}{k}}$$

If the overall similarity score is greater than a pre-defined threshold, the algorithm considers the two records matching and puts them into the same cluster. This algorithm also assumes a transitive matching relation. That is, if record *A* matches record *B* that matches *C*, the three records belong to the same cluster.

The threshold value can affect the accuracy of matching decisions (Wang et al., 2004). A large threshold value often yields a low false positive rate as well as a high false negative rate. On the other hand, a small threshold value may lower the false negative rate at the cost of a high false positive rate.

Within-cluster Pair-wise Comparison

After the clustering algorithm generates the clusters, each pair of person records within a cluster is compared based on two types of similarity measures: *personal similarity* and *social similarity*. The personal similarity measures are already presented in the clustering subsection. The social similarity measures calculate the Euclidean distances between the social features of two persons.

Personal role similarity. The role-based similarity between two persons, *i* and *j*, is based on the Euclidean distance (denoted by $||\cdot||$) between the personal role features of *i* and *j*:

$$s_r(i, j) = 1 - \| f_r(i) - f_r(j) \|.$$

Thus, the more similar two persons' personal feature values are, the more likely it is that they are the same person. However, this measure must be used with caution in two situations: *different-persons-perfect-similarity* and *same-person-small-similarity*. In the first situation, two different people may have identical values on their personal role features. For example, they might each have committed one crime and been identified as a suspect. Their personal feature values would be exactly the same, resulting in a personal role similarity of 1. In the second situation, a single person may use different identities that are associated with different roles. For example, a drug dealer may use a deceptive identity in narcotic drug crimes, while he/she also happens to be a victim in family abuse crimes in which he/she uses another identity. In this case, the personal similarity between the two personal role features will be very small even though they are the same person. The personal role similarity measure must be used together with other features such as the group role similarity.

Group role similarity. Because each of the two persons belong to a certain social group represented by their direct neighbors, the more similar the characteristics of the two groups are the more likely it is that the two persons are the same person. Like the personal role similarity, the group role similarity is defined as:

$$s_r^g(i, j) = 1 - \| f_r^g(i) - f_r^g(j) \|.$$

The group role similarity measure is also subject to the two problems of personal role similarity. First, two different persons may have exactly the same group role characteristics, causing the group similarity between them to be 1. In a clique, in which all members are fully connected with one another, the group characteristics are the same for all members. In this case, the group similarity cannot differentiate group members from each other. Second, a single person may belong to multiple social groups, each of which is associated with a specific identity. Different identities used by a single person cannot be matched successfully because of the small group similarity.

Person-associated crime type similarity and *group-associated crime type similarity* are defined similarly as:

$$s_c(i, j) = 1 - \| f_c(i) - f_c(j) \|,$$
$$s_c^g(i, j) = 1 - \| f_c^g(i) - f_c^g(j) \|.$$

Structural similarity. A criminal can be deceptive about his/her personal identity; however he/she may consistently interact with the same set of people. The more common neighbors two persons share, the more likely it is that they are the same person. Using N_i (or N_j) to denote the set of person *i*'s (or *j*'s) neighbors, the structural similarity between two persons *i* and *j* is defined as:

$$SS(i, j) = \frac{| N_i \cap N_j |}{| N_i \cup N_j |}.$$

Note that the symbol $|\cdot|$ stands for the cardinality of a set. Figure 3(a) illustrates an example in which persons 1 and 2 share the same neighbors, 4 and 7. The structural similarity between nodes 1 and 2 is thus 0.22 (= 2/9). Again, the structural similarity cannot differentiate group members in a clique because they all have the same set of neighbors.

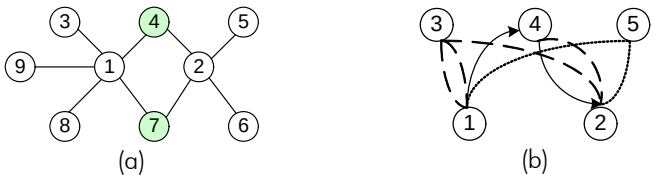

Figure 3. (a) Structural Similarity (b) Relational Similarity

Note that the structural similarity is related to the *structural equivalence* measure in social network analysis (Wasserman and Faust, 1994). Structural equivalence determines whether two persons are connected with other persons in a social network in exactly the same manner. The purpose of structural equivalence is to identify two distinct persons who can replace each other in a network, thus it is different from our goal.

Relational similarity. A link generated between two persons in the co-occurrence analysis may result from multiple incidents. In different incidents the two persons may have different role-based relations. The relation here refers to the role pairs (e.g., suspect-suspect, suspect-victim) corresponding to the two persons and is different from a co-occurrence link. One single link can consist of multiple relations. In Figure 3(b), for example, there are two relations but only one link between persons 1 and 3. Two criminals may have a suspect-suspect relation in a drug sale incident, while the conflict between them may result in an assault crime in which their relation becomes arrestee-victim.

In Figure 3(b) each arc represents a role-based relation. The dashed arc represents a suspect-suspect relation, and the dotted arc suspect-arrestee. A relation can be directional. The arrowed solid arc in Figure 3(b) means that the person at the origin of the arc is a suspect and the one at the arc head is the victim. It can be seen that both persons 1 and 2 co-occur with their neighbors in multiple incidents with multiple types of relations.

We use N_{ij} to denote the common neighbors of two persons i and j. Their relational similarity can be defined as the proportion of matched relations among i's and j's relations with their common neighbors. Again, the symbol $|\cdot|$ denotes set cardinality.

$$RS(i,j) = \frac{|\bigcap(i\text{'s relations with } N_{ij},\ j\text{'s relations with } N_{ij})|}{\text{Max}(|\ i\text{'s relations with } N_{ij}\ |,|\ j\text{'s relations with } N_{ij}\ |)}.$$

In this example, persons 3, 4 and 5 are the common neighbors of 1 and 2. Persons 1 and 2 have one matched relation with 3, one with 5, and none with 4. Persons 1 and 2 each have four relations with their common neighbors. Therefore, the relational similarity between persons 1 and 2 is 0.5 (= 2/4).

Classification

The final matching decision is made in this process. After pair-wise comparison, each identity pair in the same cluster is associated with a vector of similarity values. We treat identity matching as a binary classification problem. Specifically, based on these values a classifier can be used to further determine whether a pair of persons actually refers to the same person. Such a classifier for identity matching must be trained based on labeled data. The training set contains identity pairs with their similarity values and a class label, *match* or *non-match*. The labels are assigned to identity pairs based on either domain expert input or "gold standard criteria" created by domain experts. A classifier can be trained and a decision model can be constructed using different classification algorithms, such as decision trees (Quinlan, 1986; Quinlan, 1993), Bayesian classifier (Langley and Sage, 1994), Support Vector Machines (Cristianini and Shawe-Taylor, 2000), and so on. A test set is used to evaluate the performance of the trained classifier by comparing the predicted outcomes with the assigned labels.

Design Evaluation: Identity Matching in Illegal Drug Crimes

In a design process the evaluation is as crucial as the construction of the design artifact (Hevner et al., 2004; March and Smith, 1995). The evaluation is intended to determine if progress has been made by examining the effectiveness of the artifact. Hevner et al. (2004) listed five design evaluation methods (Guideline 3): observational, analytical, experimental, testing, and descriptive. In this research we conducted experiments on real data about illegal drug crimes in order to evaluate the effectiveness of our method for matching criminal identities. These experiments were intended to assess the utility of the method in the particular context of drug crimes—a specific complex problem—rather than provide a theoretical proposition that can be used to understand and explain the reality in the general context of complex problem solving. We begin with the description of the dataset and the definition of the evaluation metrics.

The Drug Crime Data

The problem of criminal identification is central to law enforcement and intelligence communities. In our experiments, we used the "Meth World" data provided by the Tucson Police Department (TPD) (Xu and Chen, 2003). This dataset contained records of criminals who committed crimes related to methamphetamines (a type of illegal narcotic drug) in Tucson, Arizona from 1983 to 2002. The TPD provided a list of 103 major criminals in the dataset and more than 1000 other criminals who were related to these major offenders. These criminals were involved in 28,645 crime incidents ranging from theft and aggravated assault to drug offenses. There were 23,701 person records associated with these crime incident records.

Evaluation Metrics

We evaluated matching performance using three metrics: *precision*, *recall*, and *F-measure*. These metrics are commonly used in research on record matching and information retrieval (Bilenko et al., 2003; Chen and Chau, 2004; Davis et al., 2005; Davis et al., 2003). Based on the similarity of two identity records, our method classified the record pair into one of the two categories: match and non-match. Compared with the truth, the classification (prediction) could have four outcomes as shown in Table 1.

Table 1: Prediction Outcome Matrix

Truth / Prediction	Identities of the same person	Identities of different persons
Match	True Positive (TP)	False Positive (FP)
Non-match	False Negative (FN)	True Negative (TN)

The three performance metrics are defined as follows:

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

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

$$F\text{-}measure = \frac{2 \times precision \times recall}{precision + recall}.$$

Precision measures correctly predicted matching pairs out of all pairs that are classified as matches. Recall measures correctly predicted matching pairs out of all truly matching pairs. Because the two metrics trade off against each other (one improves at the cost of the other), F-measure provides a balanced single score that calculates the weighted harmonic mean of precision and recall (Bilenko et al., 2003; Chen and Chau, 2004).

In our experiments we only computed a static set of precision, recall, and F-measure. One may be interested in checking precision ratings at different recall levels. A Precision-Recall graph (PR) can be drawn to serve this purpose. However, the PR is potentially sensitive to skewed class distributions (Fawcett, 2006). A Receiver Operating Characteristics (ROC) graph is considered better than PR when dealing with datasets with skewed class distributions (Fawcett, 2006). A ROC graph is a two-dimensional graph in which ROC-True-Positive rate is plotted on the y-axis and ROC-False-Positive rate is plotted on the x-axis. They are defined as follows:

$$\text{ROC} - \text{True} - \text{Positive} = \frac{TP}{TP + FN},$$

$$\text{ROC} - \text{False} - \text{Positive} = \frac{FP}{FP + TN}.$$

When comparing the performance of two classifiers using ROC curves, a classifier is optimal if and only if its ROC curve lies on the convex hull of the set of points in ROC space. An example is illustrated in Figure 4. In our experiments we provided ROC curves as a validation of our precision-recall performance metrics.

Hypotheses

In order to rigorously demonstrate the efficacy of our proposed method for identity matching, we select a widely used validation method, empirical hypothesis testing, in our evaluation study (Zelkowitz and Wallace, 1998). We expect that the incorporation of social contextual information would improve the performance of identity matching with data of various quality levels. Particularly, one of the biggest data quality issues in law enforcement databases is missing values. The TPD database, for example, maintains about 1.3 million person records. Each record identifies a person by a set of attributes. Except for the name attribute, whose value is mandatory, all other attributes can be potentially missing. We found that 11percent of DOB values and 28percent of address values were missing in our data and we evaluated the effectiveness of

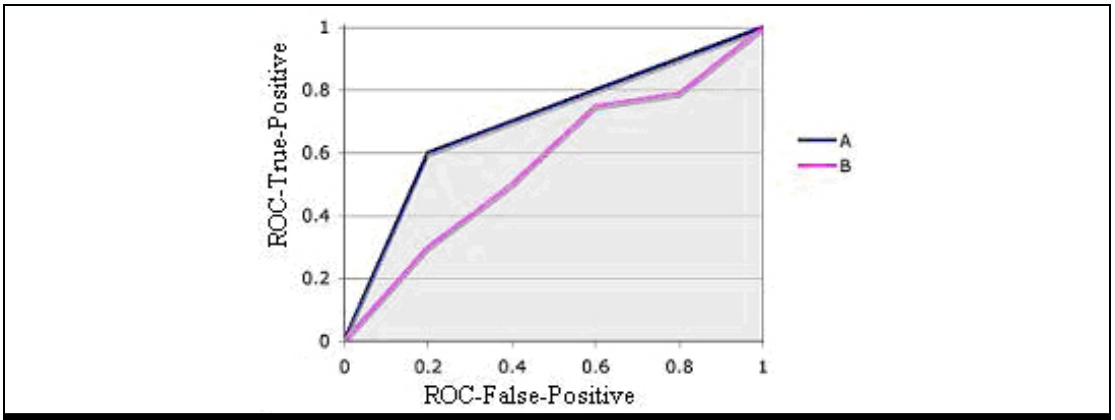

Figure 4. An illustrative example of a ROC graph. Curve A is optimal compared to curve B because it lies on the convex hull (the shaded area) of all the points in ROC space.

social contextual information on identity matching where missing values were present. An extreme case of the missing value problem would be that two identity records being compared both have name values only. When considering personal features alone, it is very unreliable to determine matching identities merely based on names. We expect that the consideration of social features would become more and more critical as the extent of missing personal features increases. Our hypotheses are stated as follows.

H1: The incorporation of social contextual features improves the performance of the identity matching technique when data are complete.

- H1.1: The use of social features in addition to personal features improves the *precision* of identity matching.
- H1.2: The use of social features in addition to personal features improves the *recall* of identity matching.
- H1.3: The use of social features in addition to personal features improves the *F-measure* of identity matching.

H2: The social contextual information improves the performance of identity matching for datasets with missing values.

- H2.1: The higher the data incompleteness level is, the more social features will improve the *precision* of identity matching.
- H2.2: The higher the data incompleteness level is, the more social features will improve the *recall* of identity matching.
- H2.3: The higher the data incompleteness level is, the more social features will improve the *F-measure* of identity matching.

We would like to make it clear that although we presented the evaluation in the form of "hypotheses," they are different from those found in traditional positivist natural science research. The hypotheses do not contain variables used to "explain how and why things are," but metrics that measure the performance of the proposed method in order to assess the utility in "serv[ing] human purposes" (March and Smith, 1995, p. 253).

Experiments

In order to investigate the effectiveness of the proposed method with missing values, we artificially constructed incomplete datasets from the complete dataset by randomly choosing a percentage of person records and removing their DOB or address values. We did not remove name values because this attribute was mandatory. The SSN was used as the gold standard for locating truly matching identities. We varied the percentage of records with missing values from 10percent to 50percent in increments of 10percent. Therefore, we had six datasets with various levels of missing values, including the complete dataset.

For each dataset, we first used the adaptive detection algorithm to pre-cluster person records based on personal features (first name, last name, DOB, and address). Two parameters need to be determined for this algorithm: window size and threshold. The window size w specifies the number of nearby records to be compared during the clustering. It usually increases as the size of the dataset increases. However, a larger window size will result in more comparisons and slow down the process. We set $w = 4$, a number proposed in a previous study with a comparable dataset (Wang et al., 2006). A threshold also needs to be determined so that two identity records being compared can be put into the same cluster if their similarity score is greater than the threshold. A small threshold value may lead to large clusters that contain many false positive matches. In our experiments we clustered the identity records in each dataset using five arbitrarily chosen threshold values: 0.75, 0.80, 0.85, 0.90, and 0.95.

Journal of the Association for Information Systems

After clustering, we considered person records within the same cluster to be candidates for matching identities, while we considered records from different clusters to be irrelevant. During within-cluster pair-wise comparison, we compared every pair of identity records in the same cluster using personal feature similarities as well as the six social similarity measures defined earlier. During the classification phase, we employed a decision tree algorithm called J48 to learn the classification model. Implementation of J48 was provided by the WEKA data mining package (Witten and Frank, 2005). We chose a decision tree classifier because of its good interpretability, whichcan help analyze the distinguishing power of different features in identity matching decisions.

We compared the performance of identity matching using both personal and social features ($F_p + F_s$) with that using personal features (F_p) alone. For each incomplete dataset, we conducted a standard 10-fold cross validation to compute the performance metrics. A statistical *t*-test was conducted for each 10-fold cross validation. We did not compare our method with other existing techniques such as the data association method (Brown and Hagen, 2002), which only suggests ways for calculating similarity between records and provides no model or rules for matching decisions. More importantly, our evaluation study was intended to examine the discriminating power of social features for identity matching, rather than to compare different techniques or algorithms. Therefore, in the experiment we compared the performance for using both personal and social features with that for using personal features only, which could be regarded as a benchmark.

Results

The Effectiveness of Social Contextual Features (H1)

Table 2 reports the average precision, recall, and F-measure ratings for identity matching using *Fp* alone and that using *Fp* and *Fs*. The better performance for each metric is highlighted. Compared with the performance using personal features alone, social contextual features significantly improved recall from 53.56percent to 66.60percent and overall F-measure from 59.52percent to 68.39percent ($p < 0.001$). Therefore, Hypotheses H1.2 and H1.3 were both supported. H1.1 was not supported because precision decreased when considering social features. Our results showed that social contextual features significantly reduced false negative matches and caused slightly higher false positive rates. The increase in F-measure showed that the overall effectiveness of identity matching was improved when social features were used along with the personal features.

Table 2: Average Performance Differences between Fp and (Fp+Fs)

Features	Precision	Recall	F-measure
Fp	**78.58%****	53.56%	59.52%
Fp + Fs	71.51%	**66.60%****	**68.39%****

Notes: *$p < 0.1$, **$p < 0.001$.

The Effects of Social Contextual Features with Incomplete Datasets (H2)

We examined the effectiveness of social contextual features for identity matching when datasets had different levels of missing values. We conducted experiments using the five artificially constructed incomplete datasets. For each performance metric, we calculated the performance differences between *Fp* and *Fp+Fs* at the five threshold levels and charted the differences against the percentage of incomplete records in a dataset.

Figure 5 (a)-(e) presents the trends in performance changes for the five threshold values (0.75, 0.80, 0.85, 0.90, and 0.95). These curves show consistent patterns. First, the curves of recall and F-measure are above the horizontal axis, while the curves of precision are below the axis. A curve above the horizontal axis means positive value changes (i.e., improved ratings). That confirmed our findings regarding H1, i.e., the incorporation of social contextual features improved recall and F-measure with slightly decreased precision ratings. Second, the extent of the improvement in recall and F-measure increased when more personal feature values were missing. The distance between a point on a curve and the *x*-axis indicates the extent of performance change. All the curves show increasing recall and F-measure as the percentage of missing values increases.

To test Hypothesis H2, we did a regression analysis on the effect of the percentage of incomplete records on the performance change. The *p*-value for the regression coefficient indicates whether or not data incompleteness significantly affects the performance changes brought about by the social contextual features. Table 3 presents the coefficients of hypothesis testing for H2.1~2.3. H2.1 was supported at $p = 0.1$ except for the threshold of 0.95. Both H2.2 and H2.3 were supported, meaning the improvements in recall and F-measure were significantly related to data incompleteness. Overall, the experiments showed that at different threshold levels, as the dataset had more missing values in personal identity attributes, the performance improvement (in recall and F-measure) brought about by the social contextual features increased significantly.

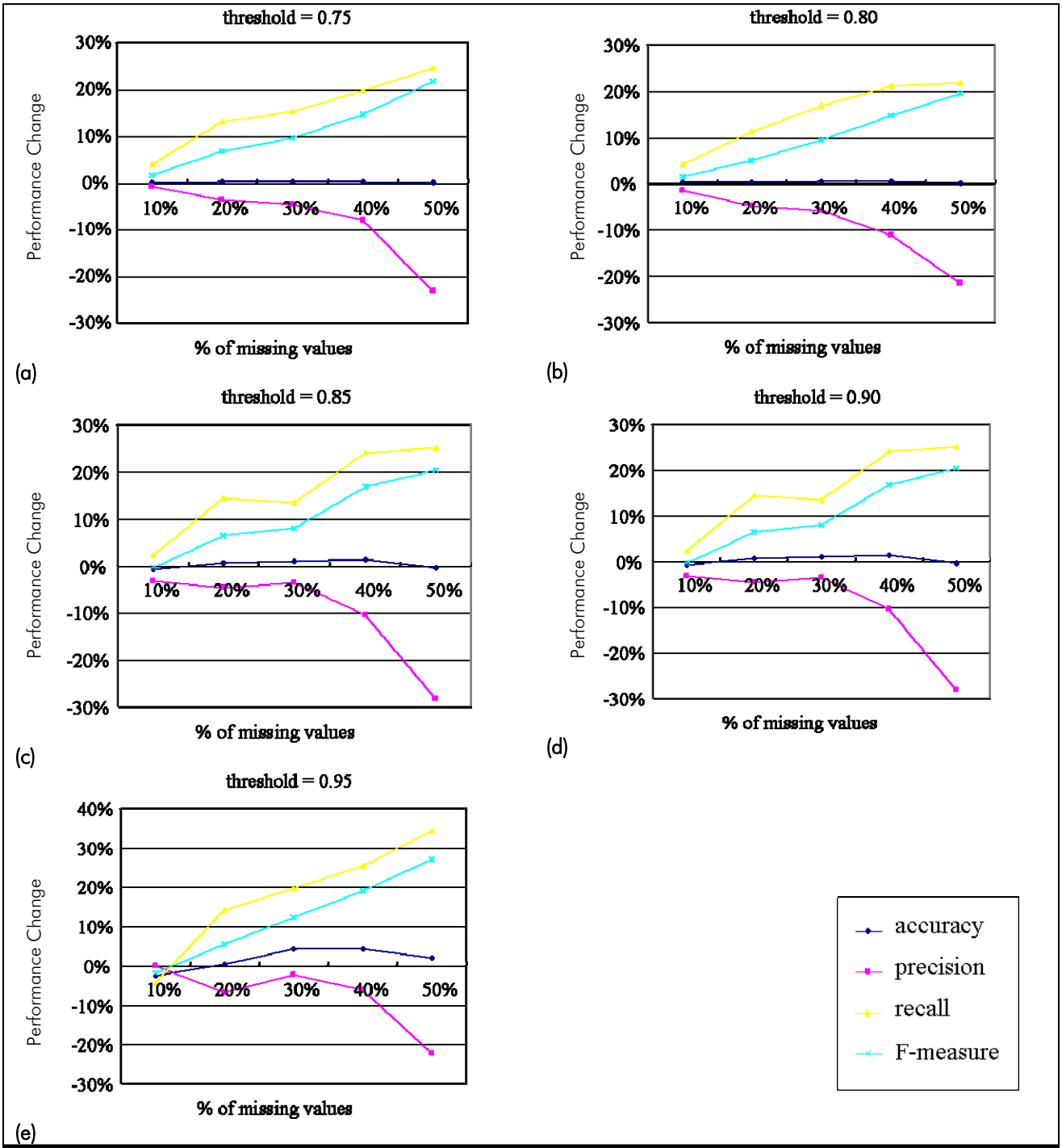

Figure 5. The Effect of Social Contextual Features on Identity Matching with Incomplete Datasets

Table 3: Statistical Testing Results for the Effect of Data Incompleteness on the Performance Change Brought by Social Features (H2.1~2.3)

Threshold	Precision	Recall	F-measure
0.75	-0.4900*	0.4783***	0.4821***
0.80	-0.4648*	0.4501***	0.4578***
0.85	-0.5605*	0.5549**	0.5193***
0.90	-0.4191*	0.5560***	0.4995***
0.95	-0.4360	0.8840***	0.7132***

Notes: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

ROC Graphs

For each incomplete dataset, we drew a ROC graph to evaluate the performance differences across different threshold values. Figure 6 (a-e) shows the ROC graphs for the five incomplete datasets. Our ROC curves are different from typical ones that usually begin at the point (0,0) and span through the point (1,1). Most curves shown in Figure 6 are partial because of the two-step analysis. The threshold values were used in the pre-clustering phase, while the ROC-false-positive and ROC-true-positive evaluated the performance of the post-clustering classification. Overall, we noticed that identity matching using both social and personal features was optimal when 20percent or more of the personal feature values were missing. This observation confirms our experimental results in precision-recall metrics.

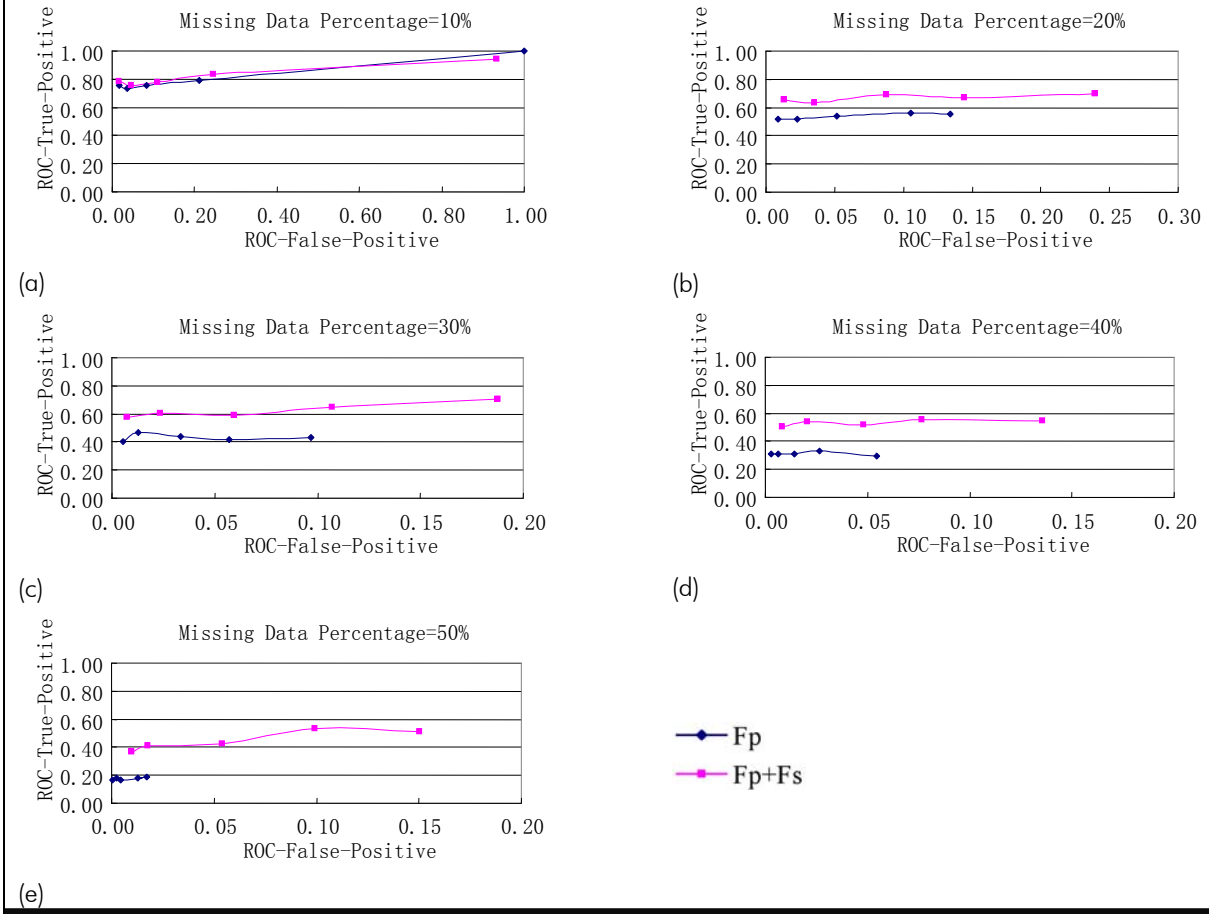

Figure 6. ROC Curves

Discussion

How Social Features Help Match Identities

Our evaluation results showed that the incorporation of social contextual features did improve the overall identity matching performance. The performance gain increased drastically when more personal features were missing. Our analysis showed that the increase in the recall rates actually contributed to the performance improvement. When the percentage of incomplete records increased, the precision rates decreased. Comparing with the outcome matrix in Table 1, we found that the decrease in precision was caused by increased false positive predictions. This means that when social contextual information was incorporated in the identity matching process, it became more likely that two different people would be considered to be the same person. As we mentioned earlier, this may result from problems such as cliques, in which different group members have exactly the same social relationships to each other. On the other hand, the reduction in false negative predictions makes it more possible to capture true matches that cannot be captured based on personal features alone.

To illustrate how social contextual information helped match identities, we examined the data and selected two cases where social features rectified the false predictions made by personal features alone. Personal feature values of the records in the two cases cannot be revealed because of the data confidentiality consideration.

The first case illustrates how social features helped to detect two matching identities when personal feature values suggested a non-matching (false negative) prediction. In this case, both record *A* and record *B* actually referred to the same person. The similarity score based only on personal features was 0.75. The classifier based on this similarity score failed to match them and considered them to be two different identities. However, the social features showed that these "two persons" had identical personal roles, always interacted with the same group of people, and were involved in the same types of crimes. Thus, with social features incorporated, these two records were classified as matching identities representing the same person.

The second case illustrates how social contextual information corrects a false positive prediction made by personal features alone. In this case, two different persons *C* and *D* happened to have identical first and last names. Their DOB and address values were also very similar. Based on personal features alone they were predicted to be the same person. When their social features were taken into consideration, however, discrepancies were found between them. Particularly, person *C* was mostly involved in assault and offense crimes, while person *D* was mainly involved in theft and drug crimes. In addition, by looking at their social relationships, most of *C*'s "neighbors" were victims in assault offenses, while *D*'s "neighbors" were often suspects or arrestees in drug-related crimes. Those disagreements in social features represented their different social behavior and, therefore, differentiated person *C* from person *D*. The similar personal feature values might result from one person intentionally concealing his identity by using the other person's identity. These two cases demonstrate the usefulness of social features in complementing personal features in identity matching.

The Larger Context of the Identity Matching Problem

The identity matching problem does not exist by itself; it should be considered in the larger context of the illegal drug problem. In this larger context, Mumford's three-stage framework provides a very good guideline for understanding and addressing this complex problem. Although our research was not directly motivated by Mumford's framework, it could well fit into the framework. First, the framework is general enough that any complex problem could be addressed through the three stages. Our research thus fits in naturally through addressing the specific facet of the illegal drug problem. Second, law enforcement and intelligence communities have long been going through these three stages to fight drug crimes (Mumford, 1999). Our research was directly motivated by the fact that these agencies often experience great difficulty matching criminal identities when they go through the third stage to investigate and reduce drug crimes. In the following paragraphs, we discuss how our research could be integrated into the framework in each stage.

Seeing the total picture. The identity matching problem is a part of the illegal drug problem. As Mumford suggested, the parts must be understood in terms of the whole (Mumford, 1998). The illegal drug problem is related to and has negative impact on many aspects of our lives. The total picture includes not only individual victims' psychological, mental, and health issues caused by drug use but also the society's social, economic, political, legislative, and security issues that result from drug use and dealing (National Drug Intelligence Center, 2007).

Figure 7 presents the various aspects of the illegal drug problem. Identity matching, as a part of the drug problem, deals primarily with the security aspect in which the goal is to fight illegal drug-related crimes such as drug trafficking, arms smuggling, assault, robbery, kidnapping, theft, etc. It also contributes to the economic aspect of the problem because criminals have to "turn their profits into a legal form of finance" by money laundering (Mumford, 1998, p. 449).

The most crucial task in fighting drug crimes is to accurately identify and capture the people who are involved in the process of drug production, transportation, distribution, sales, and money laundering. Identity matching is an inherent part of this picture because identity deception is one of the important means by which criminals protect themselves. They falsify their true identities, impersonate other individuals' identities, or use forged identity documents, hoping to mislead investigations against them. With deceptive identities, drug barons and dealers can continue to avoid being captured by law enforcement and security agencies or hide illegal financial transactions (e.g., money laundering) behind legal business activities, thereby protecting their economic profit (Mumford, 1998). Because of the potential financial loss and damaging effects that criminal identity deception may cause for victims and society, identity deception detection and matching has become one of the most important tasks in law enforcement and intelligence agencies (Brown and Hagen, 2002; Wang et al., 2004). Our research is designed to help law enforcement and intelligence agencies effectively and efficiently perform this task and better tackle drug crimes in the larger context.

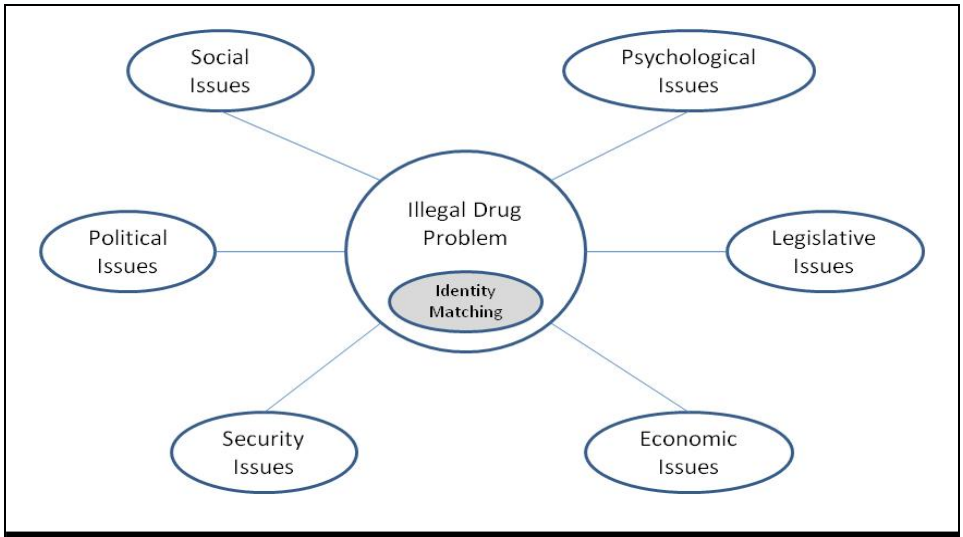

Figure 7. The Various Aspects of the Illegal Drug Problem

As our research only addresses one sub-problem in one aspect, more research must be done to address other sub-problems in other aspects. The legislative aspect, for example, deals with important questions such as "Why are drugs illegal?", "How has it arisen and why?" (Mumford, 1998, p. 454), and "Will the drug problem disappear if drugs are legalized?" There have been proposals to legalize drugs in recent years, in the hope that legalizing, taxing, and regulating the drug dealing business would solve the problems, just as the United States solved alcohol-related organized crimes such as those committed by Al Capone during the alcohol prohibition era in the 1920s (Schaffer Library of Drug Policy, 2007). However, these proposals are not likely to be accepted widely in the near future, and the illegal drug problem will continue to exist for a long time.

Developing strategies. Because the total picture of the illegal drug problem includes multiple facets, the strategies selected to tackle it should also incorporate a variety of aspects. The United States government has developed five strategies to combat the illegal drug problem (Mumford, 1999): (1) help improve the political and military abilities of nations such as Columbia and Peru; (2) enhance law enforcement and intelligence abilities to arrest and prosecute criminals; (3) closely track the movement of illegal drugs and cut their supply; (4) strengthen the U.S. border; and (5) reduce the demand for illegal drugs through educational and medical programs.

Our research fits in strategy (2) by helping enhance law enforcement and intelligence agencies' capabilities, competence, and coordination, which are the three indispensable skills for solving complex problems (Mumford, 1998; Mumford, 1999). The proposed method can be used to improve crime investigation abilities by matching criminal identities more effectively. More importantly, we believe that this method can help information sharing, collaboration, and coordination among agencies. Because drug crimes are committed by networked individuals and groups, the drug fighting forces should also be a network of personnel, agencies, and organizations. It is through such networks that people generate and exchange ideas and information. Information about drug barons, smugglers, and dealers may be scattered among different sources such as federal, state, and local law enforcement agencies, national security and intelligence organizations, customs, immigration departments, prisons, and correction communities. Coordination, collaboration, and information sharing among these agencies could help identify their common targets and assemble originally separate pieces into a big picture. These become possible only if the crime fighters have sufficient capability and competence to "search for, analyze, and synthesize relevant information [from different sources] and to relate it to past, current, and future events" (Mumford, 1998, p 451).

Additional research is especially needed in the area of international collaboration that is touched upon in strategy (1). Drug organizations and crimes have strong transnational characteristics. They spread in multiple countries and regions and cannot be completely resolved by any single government. How to support and enhance international collaboration and coordination is one of the most important questions facing nations that are determined to tackle the illegal drug problem.

Taking action. This stage is aimed at operationally tackling the problem following the strategies developed earlier. This is a stage in which information technology can play a significant role in enabling the strategies by enhancing the capability, competence, and coordination of problem solvers. Although there is not a single technology that can solve the drug problem as a whole, the problem can be broken down into sub-problems that are more manageable and solvable. As new

technologies are constantly being designed and developed, problem solvers will be in a better position to solve the complex problem.

The focus of our present study is primarily on this stage in which we seek to use information technology to aid in fighting drug crimes. Our research is designed to improve the effectiveness of criminal identity matching methods by incorporating the criminals' social contextual information in the matching process. Although the problem we have studied is only a small sub-problem in the larger context of illegal drugs, our research demonstrates how a seemingly intractable, insolvable complex problem can be partially addressed by taking one aspect of the big picture and moving forward one step at a time along the path to the solution.

Conclusions and Future Directions

Complex problems are characterized by a large number of variables that interact with each other and constantly change over time. In the context of drug crimes and other organized crimes, the complexity to a large extent comes from the criminal organization itself, which takes the form of networks of people and relationships. This implies that when trying to solve the problem of drug crimes, the problem solvers must always keep in mind that criminals are not isolated individuals but connected by various relationships. In this paper we report our research on using social contextual information to help tackle a sub-problem of drug crimes: the identity matching problem. We treated criminals as actors in social networks and utilized the information about the social associations between them to match seemingly unrelated criminal identity records. The evaluation results were encouraging and supported our expectation that combining social features with personal features could help match criminal identities more effectively. The social contextual features became increasingly useful when the data contain more missing values.

Our research focused on the third stage (namely, taking action) of Mumford's three-stage framework for solving complex problems. In this stage we followed the seven guidelines of design science research (Hevner et al., 2004). Several guidelines have been discussed in the previous sections. Here we discuss the remaining three guidelines. Our research rigor (Guideline 5) was achieved through both the construction and evaluation of the design artifact. Our method was constructed based on theories of social identity, and all the features and measures were carefully selected and defined. The method's performance was evaluated against several metrics to test the two hypotheses regarding the usefulness of social contextual features. Although our paper is primarily aimed at technical audiences, we present our research in a manner that would also be of interest to managerial audiences (Guideline 7). Managers and decision makers of organizations, for example, may focus on the discussion about how a complex problem like drug crimes can be addressed by going through the three stages suggested by Mumford (1998).

Implications for Practice and Theory

Our research has important implications for both practice and theory (Guideline 4). In practice, our method can help law enforcement personnel and intelligence agencies match people's identities more effectively. To the best of our knowledge, our method is the first identity matching technique that utilizes social contextual information. Throughout the project, we have worked closely with personnel in a city-level police department and tested our technique using real criminal data. We believe our approach of combining personal and social features can be extended to other law enforcement data or even to other applications such as anti-terrorism. With criminal identities being matched more effectively and efficiently, information sharing and collaboration across jurisdictions will become more possible and feasible. Advances in solving this sub-problem will provide greater capability and competence to solve the bigger, more complex drug and other crimes.

In terms of theoretical implications, our research instantiates and extends Mumford's idea about complex problem solving. Mumford used illegal drugs as an example of a complex problem. She illustrated the problem and suggested that one could use the three-stage problem solving framework to tackle the problem. We instantiated her idea by recognizing identity matching as a sub-problem of illegal drugs and designed a new method for addressing the problem. In addition, we used a real dataset about drug criminals and our study demonstrated that the three-stage framework does have a great value and utility in solving real problems.

Our research draws the connection between Mumford's problem solving framework and the design science paradigm, which are both about problem solving. We demonstrate that a design science study can be positioned in the three-stage framework, especially in the third stage (*taking action*). We believe that the knowledge and experience gained from building, applying, and evaluating design artifacts (i.e., constructs, models, methods, and instantiations) in this stage can provide invaluable feedback to the other two stages, especially strategy improvement. As design science is getting more attention in the information systems research community in recent years, we hope that our study brings more recognition and

appreciation of Mumford's framework, as it emphasizes not only the operational stage but also the two preceding stages that necessarily guide the design of solutions to complex problems.

Moreover, one key question raised by Mumford is "how can technology assist?" (Mumford, 1999, p. 198). By using the design science methodology under Mumford's three-stage framework, our study serves as an example of information technology's role in enabling the strategy for information sharing and collaboration between agencies in crime fighting. In Mumford's framework, information technology does not live in a vacuum but is an organic, inherent component of the complex problem solving process. Any IS researcher, when studying IS artifacts and phenomena, should think about the "big picture" and the strategy issues: What is the role of information technology in the total picture of organizations? How can we align information technology with business strategy (Henderson and Venkatraman, 1993)? How can information technology assist and enable the solution to complex problems that organizations experience?

We believe that Mumford's use of drug crimes as an example of a complex problem has another important implication for the IS community. IS research, which has been primarily focused on business-related IT phenomena, should go beyond the current scope to "both the private and public sectors, to individuals, organizations, and transnational organizations" (Baskerville and Myers, 2002, p. 6), and to the whole society. Information technology should be studied at all levels and aspects of our lives to solve various complex problems.

Limitations and Future Research

While the results of our study are encouraging, a few limitations of the study should be noted. First, our approach was only tested on one dataset, namely the dataset on illegal drug criminals in a U.S. city. Although we believe the approach can be applied to other datasets for identity matching, our design that incorporates social contextual features may not be applicable to all types of crimes and all geographic areas, due to the varying nature of crimes or cultures. For example, cybercrime, the other example of a complex problem used in Mumford's illustrations, is often committed solitarily. In such cases, social contextual features may be less prominent in the data and thus less effective.

Another limitation is that we only tackled a sub-problem of the illegal drug problem. What we did not study is the overall effectiveness of our approach on the total picture, i.e., fighting against illegal drugs. This is an inherent issue in complex problem solving. Given the large number of variables involved in the problem, it could be difficult to isolate a sub-problem and study its overall effect on the main problem. Nonetheless, based on Mumford's framework, organizations' capabilities in the problem solving areas are required for successful problem solving. The approach reported in this paper can, undoubtedly, improve such capabilities.

Moreover, although the drug problem itself is dynamic, the identity matching process we present in this study does not show much dynamic nature because we used a fixed, "static" dataset. In practice, the problem will become more dynamic because as criminals continue to carry out illegal activities, the data about them and their crimes will also change over time.

"[P]roblem solving is a difficult process that can always be improved" (Mumford, 1998, p. 457). We hope to keep improving our technique in our future research, which can be conducted in several directions. First, it is possible to incorporate more social contextual features that can be used to match criminal identities. For example, social features obtained from other data sources (e.g., data from other government entities such as Customs) can be used. It would also be possible to apply social network analysis techniques in the process. In addition, different classification models could be tested to further evaluate the performance of our approach.

Another avenue of future research would be to apply our proposed approach to identity matching to other complex problems and then evaluate its performance. Examples include customer relationships management, anti-terrorism investigation, and fraud detection. Based on the results of the present study, it will be interesting to see whether identity matching techniques and social contextual features can help improve the accuracy and performance in these applications.

In summary, complex problems are difficult to solve. Without a thorough understanding of the characteristics of the problem and well-developed strategies, the process of seeking the solution would be "similar to shooting in the dark" (Mumford, 1998, p. 456). The identity matching problem, which has been addressed by existing techniques solely from an individual perspective, becomes more solvable when viewed in the total picture of the illegal drug problem and other organized crimes. Recognizing the fact that criminals are actors in social networks of relationships, we have found new social identity features that can help tackle the problem. Although our method is still not optimal, it has been carefully thought out and designed, which we hope will lead to "a logical path to a solution" to the complex problem of crimes (Mumford, 1998, p. 456).

Acknowledgements

This research is supported in part by the following grants:
NSF Digital Government Program, "COPLINK Center: Social Network Analysis and Identity Deception Detection for Law Enforcement and Homeland Security," #0429364, 2004-2007.
NSF Digital Government Program, "COPLINK Center: Information and Knowledge Management for Law Enforcement," #9983304, July 2000 – September 2006.

We would like to thank Rudy Hirschheim and Jaana Porra, the guest editors of this special issue, and anonymous reviewers for their insightful comments that substantially enhanced the quality of the paper. Special thanks go to Hsinchun Chen of the University of Arizona for leading the COPLINK project and his support for this study. We would also like to thank the Tucson Police Department for providing us with the data for our analysis. In particular, we would like to thank Tim Petersen and Siddharth Kaza for serving as the domain experts in our evaluation studies.

References

Baskerville, R. L. and M. D. Myers (2002) "Information systems as a reference discipline," *MIS Quarterly* (26) 1, pp. 1-14.

Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar et al. (2003) "Adaptive name matching in information integration," *IEEE Intelligent Systems* (18) 5, pp. 16-23.

Brown, D. E. and S. Hagen (2002) "Data association methods with applications to law enforcement," *Decision Support Systems* (34) pp. 369-378.

Cheek, J. M. and S. R. Briggs (1982) "Self-consciousness and aspects of identity," *Journal of Research in Personality* (16) pp. 401-408.

Chen, H. and M. Chau (2004) "Web mining: Machine learning for web applications," *Annual Review of Information Science and Technology* (38) pp. 289-329.

Chen, H. and K. J. Lynch (1992) "Automatic construction of networks of concepts characterizing document databases," *IEEE Transactions on Systems, Man and Cybernetics* (22) 5, pp. 885-902.

Chen, H., J. Martinez, A. Kirchhoff, T. D. Ng et al. (1998) "Alleviating search uncertainty through concept associations: Automatic indexing, co-occurrence analysis, and parallel computing," *Journal of the American Society for Information Science* (49) 3, pp. 206-216.

Chen, H., D. Zeng, H. Atabakhsh, W. Wyzga et al. (2003) "COPLINK managing law enforcement data and knowledge," *Communications of the ACM* (46) 1, pp. 28-34.

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

Cristianini, N. and J. Shawe-Taylor (2000) *An Introduction to Support Vector Machines*: Cambridge University Press.

Davis, J., I. Dutra, D. Page, and V. S. Costa. (2005) "Establishing identity equivalence in multi-relational domains." *International Conference on Intelligence Analysis, McLean, VA, 2005*.

Davis, P. T., D. K. Elson, and J. L. Klavans. (2003) "Methods for precise named entity matching in digital collections." *The Third ACM/IEEE Joint Conference on Digital Libraries (JCDL), Houston, TX, 2003*.

Deaux, K. and D. Martin (2003) "Interpersonal networks and social categories: Specifying levels of context in identity processes," *Social Psychology Quarterly* (66) 2, pp. 101-117.

Dey, D., S. Sarkar, and P. De (2002) "A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases," *IEEE Transactions on Knowledge and Data Engineering* (14) 3, pp. 567-582.

Fawcett, T. (2006) "An introduction to ROC analysis," *Pattern Recognition Letters* (27) 8, pp. 861-874.

Funke, J. (1991) "Solving complex problems: Exploration and control of complex systems," in R. J. Sternberg and P. A. Frensch (Eds.) *Complex Problem Solving: Principles and Mechanisms*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 185-222.

Gray, W. D. (2002) "Simulated task environments: The role of high fidelity simulators, scaled worlds, synthetic environments, and laboratory tasks in basic and applied cognitive research," *Cognitive Science Quarterly* (2) 2, pp. 205-227.

Hauck, R., R. Sewell, D. T. Ng, and H. Chen (2001) "Concept-based searching and browsing: A geoscience experiment," *Journal of Information Science* (27) 4, pp. 199-210.

Hauck, R. V., H. Atabakhsh, P. Ongvasith, H. Gupta et al. (2002) "Using coplink to analyze criminal-justice data," *IEEE Computer* (35) 3, pp. 30-37.

Henderson, J. and N. Venkatraman (1993) "Strategic alignment: Leveraging information technology for transforming organizations," *IBM Systems Journal* (32) 1, pp. 4-16.

Hevner, A. R., S. T. March, J. Park, and S. Ram (2004) "Design science in Information Systems research," *MIS Quarterly* (28) 1, pp. 75-105.

Ianni, F. A. J. and E. Reuss-Ianni (1990) *Network Analysis*. California, USA: Palmer Enterprises.

Jonas, J. (2006) "Identity resolution: 23 years of practical experience and observations at scale." Chicago, IL, USA: ACM Press.

Langley, P. and S. Sage. (1994) "Induction of selective Bayesian classifiers." *the 10th conference of Uncertainty Artificial Intelligence, Seattle, WA, 1994*, pp. 399-406.

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

March, S. T. and G. F. Smith (1995) "Design and natural science research on information technology," *Decision Support Systems* (15) 4, pp. 251-266.

Marshall, B., S. Kaza, J. Xu, H. Atabakhsh et al. (2004) "Cross-jurisdictional criminal activity networks to support border and transportation security." *the 7th Annual IEEE Conference on Intelligent Transportation Systems (ITSC 2004), Washington, D.C., 2004*.

Mumford, E. (1998) "Problems, knowledge, solutions: Solving complex problems." *International Conference on Information Systems, Helsinki, Finland, 1998*.

Mumford, E. (1999) *Dangerous Decisions - Problem Solving in Tomorrow's World*. New York: Kluwer Academic/Plenum Publishers.

National Drug Intelligence Center (2007) "National Drug Threat Assessment 2007," http://www.usdoj.gov/ndic/pubs21/21137/index.htm (May 11, 2007).

Quesada, J., W. Kintsch, and E. Gómez Milán (2005) "Complex problem solving: A field in search of a definition?," *Theoretical issues in Ergonomics Sciences* (6) 1, pp. 5-33.

Quinlan, J. R. (1986) "Introduction of decision trees," *Machine Learning* (1) pp. 86-106.

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*. Amsterdam, the Netherlands: Morgan Kaufmann.

Redman, T. C. (1998) "The impact of poor data quality on the typical enterprises," *Communications of the ACM* (41) 3, pp. 79-82.

Schaffer Library of Drug Policy (2007) "The Lessons of Prohibition and Drug Legalization " http://www.druglibrary.org/schaffer/debate/myths/myths6.htm (September 14, 2007).

Schroeder, J., J. Xu, H. Chen, and M. Chau (2007) "Automated criminal link analysis based on domain knowledge," *Journal of the American Society for Information Science and Technology* (58) 6, pp. 842-855.

Stryker, S. and R. T. Serpe (1982) "Commitment, identity salience, and role behavior: Theory and research example," in W. Ickes and E. S. Knowles (Eds.) *Personality, Roles, and Social Behavior*, New York: Springer-Verlap, pp. 199-218.

Tajfel, H. and J. C. Turner (1986) "The social identity theory of inter-group behavior," in S. Worchel and L. W. Austin (Eds.) *Psychology of Intergroup Relations*, Chicago: Nelson-Hall.

Turner, J. C. (1999) "Some current issues in research on social identity and self-categorization theories," in N. Ellemers, R. Spears, and B. Doosje (Eds.) *Social Identity: Context, Commitment, Content*, Blackwell: Oxford.

Wang, G., H. Chen, and H. Atabakhsh (2004) "Automatically detecting deceptive criminal identities," *Communications of the ACM* (47) 3, pp. 71-76.

Wang, G. A., H. Chen, J. J. Xu, and H. Atabakhsh (2006) "Automatically detecting criminal identity deception: An adaptive detection algorithm," *IEEE Transactions on Systems Man and Cybernetics (Part A-Systems and Humans)* (36) 5, pp. 988-999.

Wasserman, S. and K. Faust (1994) *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Witten, I. H. and E. Frank (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2 edition. San Francisco: Morgan Kaufmann.

Xu, J. and H. Chen. (2003) "Untangling criminal networks: A case study." *the 1st NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03), Tucson, AZ, 2003*, pp. 232-248 LNCS 2665.

Zelkowitz, M. and D. Wallace (1998) "Experimental models for validating technology," *IEEE Computer* (31) 5, pp. 23-31.

About the Authors

Jennifer Xu is an Assistant Professor of Computer Information Systems at Bentley College. She received her PhD in Management Information Systems from the University of Arizona. Her research interests include data mining, Web mining, social network analysis, knowledge management, and information visualization. Her work has appeared in several IS journals and conferences including *ACM Transactions on Information Systems*, *Journal of the American Society for Information Science and Technology*, *IEEE Transactions on Systems, Man, and Cybernetics*, *Communications of the ACM*, *Decision Support Systems*, *International Journal of Human-Computer Studies*, and *International Conference on Information Systems*.

G. Alan Wang is an Assistant Professor of Business Information Technology at Virginia Tech, Blacksburg, VA. He received his Ph.D. degree in Management Information Systems from the University of Arizona in 2006, and his Master's degree in Industrial Engineering from Louisiana State University in 2001. His research interests include heterogeneous data management, data cleansing, data mining and knowledge discovery, and decision support systems. He has published in *Communications of the ACM*, *IEEE Transactions on Systems, Man, and Cybernetics*, *IEEE Computer*, *Group Decision and Negotiation*, and *Journal of the American Society for Information Science and Technology*.

Jiexun Li is an Assistant Professor in the College of Information Science and Technology at Drexel University, Philadelphia, PA. He received his Ph.D. degree in Management Information Systems from the University of Arizona in 2007 and his Master's degree in Management from Tsinghua University, Beijing, China in 2002. His research focuses on data/text mining and machine learning for knowledge discovery. His research in knowledge discovery has covered various application areas such as business, bioinformatics, and security. He has published in *Communications of ACM*, *Journal of the American Society for Information Science and Technology*, *Bioinformatics*, *IEEE Transactions on Information Technology in Biomedicine*, and *Decision Support Systems*.

Michael Chau is an Assistant Professor in the School of Business at the University of Hong Kong. He received a Ph.D. degree in Management Information Systems from the University of Arizona and a bachelor degree in Computer Science and Information Systems from the University of Hong Kong. His current research interests include information retrieval, Web mining, data mining, knowledge management, electronic commerce, security informatics, and intelligence agents. He has published more than 60 research articles in various journals and conferences including *IEEE Computer*, *ACM Transactions on Information Systems*, *Journal of the America Society for Information Science and Technology*, *Annual Review of Information Science and Technology*, *Decision Support Systems*, *International Journal of Human-Computer Studies*, *Communications of the ACM*, and *International Conference on Information Systems*. More information can be found at http://www.business.hku.hk/~mchau/.

Journal of the Association for Information Systems
ISSN: 1536-9323

Editor
Kalle Lyytinen
Case Western Reserve University, USA