Automated Criminal Link Analysis Based on Domain Knowledge

Jennifer Schroeder
Tucson Police Department, 270 S. Stone Avenue, Tucson, AZ 85701.
E-mail: Jenny.Schroeder@tucsonaz.gov

Jennifer Xu
Computer Information Systems Department, Bentley College, 175 Forest Street, Waltham, MA 02452.
E-mail: jxu@bentley.edu

Hsinchun Chen
Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721.
E-mail: hchen@eller.arizona.edu

Michael Chau
School of Business, The University of Hong Kong, Pokfulam, Hong Kong. E-mail: mchau@business.hku.hk

Link (association) analysis has been used in the criminal justice domain to search large datasets for associations between crime entities in order to facilitate crime investigations. However, link analysis still faces many challenging problems, such as information overload, high search complexity, and heavy reliance on domain knowledge. To address these challenges, this article proposes several techniques for automated, effective, and efficient link analysis. These techniques include the co-occurrence analysis, the shortest path algorithm, and a heuristic approach to identifying associations and determining their importance. We developed a prototype system called CrimeLink Explorer based on the proposed techniques. Results of a user study with 10 crime investigators from the Tucson Police Department showed that our system could help subjects conduct link analysis more efficiently than traditional single-level link analysis tools. Moreover, subjects believed that association paths found based on the heuristic approach were more accurate than those found based solely on the co-occurrence analysis and that the automated link analysis system would be of great help in crime investigations.

Introduction

Criminal justice is an application domain where information science and technology play an important role in facilitating the investigation of various crimes and illegal activities (Chen 2005; Strickland & Hunt 2005). Link analysis tool is such a kind of information system that law enforcement and intelligence agencies throughout the world have long used for crime investigation. Unlike link analysis in the Web computing domain—where the purpose is to analyze the hyperlink structure between Web pages to identify authoritative and hub pages (Brin & Page, 1998; Kleinberg 1999)—link analysis in the criminal justice domain refers to the identification, analysis, and visualization of relationships and associations between crime entities (e.g., persons, organizations, vehicles, locations, and criminal incidents; Harper & Harris, 1975; Mena, 2003). By establishing association paths linking known entities such as the suspect and the victim in a crime, link analysis may provide information about motives and help uncover investigative leads. For example, the Federal Bureau of Investigation (FBI) used link analysis in the investigation of the Oklahoma City Bombing case and the Unabomber case to look for criminal associations and to identify suspects. The Department of the Treasury of the United States analyzed the associations between individuals, bank accounts, and financial transactions to detect money-laundering activities (Goldberg & Senator, 1998; Goldberg & Wong, 1998).

Although it has been more than 30 years since it was introduced in 1975 (Harper & Harris, 1975), link analysis remains a challenging task and, to a large extent, a manual process for several reasons. First, the "information overload" problem (Blair, 1985) makes identifying and searching for crime associations very time-consuming. To correlate known entities in a crime incident, a crime investigator must

Received February 14, 2006; revised May 24, 2006; accepted June 10, 2006

© 2007 Wiley Periodicals, Inc. • Published online 21 February 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20552

manually search for associations by examining a large number of documents. The documents may range from structured database records of crime incidents to unstructured report narratives. The process is similar to a breadth-first search in which a search tree rooted at one of the known entities is expanded level-by-level down to other known entities. For example, to find an association path between two known entities *A* and *B*, a crime investigator first retrieves all documents containing one known entity, *A*, and looks for other entities that appear in these documents that contain *A*. These newly discovered entities are linked to *A* and the search tree is expanded to level one. If level one does not contain entity *B*, the investigator must retrieve and read more documents to expand the search tree by searching for entities associated with the level one entities. The search tree will be progressively expanded until a path connecting *A* and *B* is found. Each expansion in this process entails the investigator examining one or more documents and thus consumes much investigative time and effort.

Second, high branching factors (the number of direct associations an entity has) can increase the complexity of association path search dramatically during link analysis (Jensen, 1998). A high branching factor can lead to a large number of associations that need to be evaluated when the crime entities of interest are not directly associated. In a breadth-first search of depth four; for instance, an average branching factor of seven can result in 2,401 associations that need to be evaluated. In practice, the branching factors can be much higher for criminals who have repeated police contacts and arrests. These criminals have committed many crimes involving many people and have a large number of associations. The branching factor can be further inflated if associations with many entity types (e.g., addresses, organizations, property, or vehicles) are considered.

Moreover, the purpose of association path search is to find paths that contain important links between crime entities and can help uncover investigative leads. However, paths found using a breadth-first search may not necessarily be useful. The paths may contain trivial, unimportant links or too many intermediate links. A simple breadth-first search cannot guarantee that shorter paths containing important associations can be found.

Finally, link analysis relies heavily on crime investigators' domain knowledge and experience, making it difficult to automate. Specifically, investigators must be able to determine whether an association between two crime entities is important for uncovering investigative leads. By reading incident report narratives, an investigator often can tell whether the association in question actually exists and how strong the association is. However, structured crime incident reports provide only limited information about an incident, such as the time, location, persons, and crime type, and do not indicate explicitly the existence of an association, let alone the strength of the association. With limited information, investigators' domain knowledge and experience play a key role in judging the importance of an association. For example, whether two persons who are involved in the same crime have a strong relationship can depend on the type of crime. In homicide crimes, the suspect and the victim often know each other well or are at least acquaintances. In contrast, the suspect and victim in burglary cases often are not known to each other. Such heuristics are tacit knowledge that resides in crime investigators' minds and is difficult to model and incorporate into automated systems.

Although some link analysis software packages are available, most of them do not help identify, search, and analyze associations beyond simple visualization of crime associations. Some tools facilitate only single-level association searches, finding only directly related entities. Because of these challenges, link analysis is considered an effective but costly analysis method and is used only in the investigation of high profile crimes. Automated, effective, and efficient link analysis techniques are needed to address the challenges and to assist crime investigation (McAndrew, 1999; Sparrow, 1991).

The goal of this article is to propose several techniques for automated link analysis: the co-occurrence analysis approach (Chen & Lynch, 1992) and a heuristic approach for the identification of associations between crime entities, and a shortest path algorithm (Dijkstra, 1959; Helgason, Kennington, & Stewart, 1993) for association path search. In particular, the heuristic approach helps incorporate crime investigators' domain knowledge into a link analysis system for judging association strength automatically.

The remainder of the article is organized as follows. We review prior literature in the Literature Review section and discuss system design in the System Design section. We present our research questions in the Research Questions section and results of a system evaluation study conducted at the Tucson Police Department (TPD) in System Evaluation section. The Conclusions and Future Work section concludes the article and suggests some future directions.

Literature Review

In this section, we review prior work in association identification, knowledge engineering, and association path search. We also provide a brief review of association visualization functionality in existing link analysis tools and systems.

Association Identification

Identifying crime entity associations from structured or unstructured documents is a bottleneck of link analysis. Structured documents such as crime incident records, bank accounts, and financial transaction records usually do not contain explicit information about entity associations. Unstructured documents such as police report narratives may contain various entities and implicit association information that are difficult to extract (Li & Yang, 2005; Wu & Pottenger, 2005). Prior research has proposed some approaches to help address this problem. These approaches can be roughly divided into four categories: heuristic-based, template-based, similarity-based, and statistical approaches.

Heuristic-based approaches use decision rules that human investigators employ to find associations between crime entities. For example, the FinCEN system at the U.S. Department of the Treasury forms a transactional association between two persons if one person deposits to a bank account owned by the other person (Goldberg & Senator, 1998; Goldberg & Wong, 1998). To use this approach, a large amount of knowledge engineering effort must be invested to acquire expert knowledge and heuristics. At present, this approach is used only for structured data.

Lee (1998) proposed an approach to extract association information from unstructured, textual police documents by matching phrases or sentences to predefined templates. This approach uses Natural Language Processing (NLP) techniques to first extract from a textual document all entity types (e.g., persons, properties, locations, date, and time). It then compares the phrases or sentences containing the entities to a collection of handcrafted patterns. For example, "Smith owns a Toyota Camry" can be matched to a template "<person><own><property>," which specifies an entity-entity association. Because this approach depends entirely on handcrafted language rules and patterns, it is difficult to scale up and apply to new documents without matching templates.

In the similarity-based approach, the similarity (and dissimilarity) between entities and cases are used to identify data associations in law enforcement databases (Brown & Hagen, 2002; Lin & Brown, 2003). For example, a similarity score between two records can be calculated based on attributes such as hair color and body height of the suspects. The weight of each attribute can be dynamically adjusted. This method has been shown to achieve accuracy comparable to crime analysts with significant time required.

The statistical approach, such as the co-occurrence analysis approach, identifies associations between entities based on lexical statistics. The co-occurrence analysis approach was originally designed for generating thesauri from textual documents automatically by measuring the frequency that two phrases appear in the same documents (Chen & Lynch, 1992). Assuming that two entities appearing in the same documents may have an association, a nonzero co-occurrence weight can indicate the existence of an association. In addition, the higher a co-occurrence weight, the more likely it is that the two entities involved have a strong association. Because this statistical approach can process a large number of documents automatically, it can address the information overload problem fairly well and has been used to extract crime entity associations from both structured (Chen et al., 2004; Hauck, Atabakhsh, Ongvasith, Gupta, & Chen, 2002) and unstructured documents (Baldwin & Bagga, 1998; Chau, Xu, & Chen, 2002; Chen et al. 2004; Xu & Chen, 2004).

Identifying previously unknown associations among seemingly unrelated concepts, entities, and literatures from documents is not new to the information retrieval research community. Based on the analysis of MEDLINE literature, Swanson (1986) uncovered previously unknown linkages between fish oil and blood viscosity, and between blood viscosity and a disease called Raynaud's syndrome. This discovery led to a testable hypothesis that fish oil reduces blood viscosity, thereby helping to alleviate Raynaud's syndrome. Such literature-based discovery has later been partly automated by using lexical statistics (e.g., word frequency and co-occurrence weight) to guide the identification of hidden connections among medial literatures (Lindsay & Gordon, 1999; van der Eijk, van Mulligen, Kors, Mons, & van den Berg, 2004) and the search for new applications of existing technologies or solutions (Gordon, Lindsay, & Fan, 2002).

However, a drawback of automatic methods such as the similarity-based and statistical approaches is that they often do not take into account the nature of a relationship. Two persons who do not appear together frequently in the same crimes may actually be family members and have a strong relationship. Domain experts often make decisions about the importance of associations using a number of heuristics. Modeling and incorporating the heuristics into automated link analysis systems, however, requires knowledge engineering.

Knowledge Engineering

Determining the importance of associations between crime entities is highly dependent on the domain knowledge and experience of crime investigators. The domain knowledge includes not only *what* factors can be used to make a judgment but also *how* to judge based on the factors (Wildemuth, 2004). In link analysis, the approaches for incorporating expert knowledge have been primarily ad-hoc. As reviewed earlier, Goldberg and Senator (1998) used a heuristic-based approach in the FinCEN system to form associations between individuals who had related transactions or shared bank accounts. These heuristics were used by investigators to manually uncover associations but were not really incorporated into the system for automated link analysis. In cases with large datasets, investigators still face the problems of information overload and high search complexity.

How to model and incorporate domain experts' tacit knowledge into automated systems has been studied extensively in the knowledge engineering discipline. Two indispensable steps of knowledge engineering are the construction of a knowledge base and the development of an inference engine (Martin & Oxman, 1988). In addition to facts, ontologies, and concepts defined in a specific domain, a knowledge base contains important decision rules and problem solving methods that domain experts use to make decisions or judgments. The inference engine consults these methods while making decisions automatically.

The decision rules or heuristics are often represented as *if-then* statements (Martin & Oxman, 1988). To acquire these decision rules, a knowledge engineer must interview or observe extensively how experts make decisions. In general, knowledge acquisition includes the identification of predictive factors or variables that can affect the outcome variable, and the collection of rules that govern the determination of outcomes based on the values of the predictive variables. Many existing expert systems are built based on decision rules obtained from domain experts. Examples include factory scheduling (Fox & Smith, 1984), telephone switch maintenance (Goyal, 1985), and disease diagnosis (Shortliffe, 1976). However, knowledge acquisition is often time-consuming and difficult. A large body of research has proposed various methods to obtain decision rules automatically from data rather than consulting domain experts. Many classification methods such as ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), and neural network classifiers (Wilson & Sharda, 1994) have been applied in this line of research (Badiru & Cheung, 2002).

In situations where experts must deal with uncertainty, a rule may be associated with a confidence factor or a probability. One of the *reasoning-with-uncertainty* methods is Bayesian belief network (Heckerman, 1999). It has been found that the decision processes of experts in some situations, such as auditors' assessment of a bank's financial health, can be modeled as a belief network (Sarkar & Sriram, 2001). A belief network is a probability network with a node representing an outcome variable or a predictive variable that can affect the outcome. Links between these variables specify the dependency relationships (Heckerman, 1999). For example, an auditor's belief network may consist of an outcome variable, bank failure, and a set of predictive variables such as Return on Assets (ROA), Return on Equity (ROE), and other financial ratios (Sarkar & Sriram, 2001). Based on Bayesian rules and conditional independence, the probability of a specific outcome value o_i can be viewed as being dependent on the values of predictive variables (Langley & Sage, 1994). An advantage of Bayesian belief network is its ability to express experts' prior knowledge as a network associated with initial conditional probability distributions, which may later be revised based on new data (Heckerman 1999).

Association Path Search

Association paths between entities that are not directly connected may contain multiple, intermediate links. Literature-based discovery systems usually help users find such paths by progressively expanding the level of the source entity to multiple levels (Gordon et al., 2002; Lindsay & Gordon, 1999; Swanson & Smalheiser, 1997), similar to a breadth-first search. The result can be multiple paths leading from the source entity to the target entity (Das-Veves, Fox, & Yu, 2005).

In some domains such as the medical domain, longer paths containing many intermediate entities may be as important as shorter paths because they both provide the relationships between the source and the target. In the crime investigation domain, shorter association paths often are more likely to uncover investigative leads because they provide the most direct way to link two entities. Manually searching for the shortest association path may cost much time because of the information overload problem and high branching factors of entities. Some link analysis tools allow for "single-level" or direct association searches. The Watson system (Anderson, Arbetter, Benawides, & Longmore-Etheridge, 1994) can identify possible links and associations between entities by querying databases. Given a specific entity such as a person's name, Watson can automatically form a database query to search for other related records. The related records found are linked to the given entity and the result is presented in a link chart. The COPLINK Detect system (Chen et al., 2004; Hauck et al., 2002) can also find direct associations between entities if they appear in the same documents. However, neither the Watson nor COPLINK Detect system facilitates the search for association paths consisting of multiple hidden, intermediate links.

Researchers have proposed employing shortest path algorithms to find crime entity association paths of multiple levels. The Link Discovery Tool uses shortest path algorithms to search for the associations between two individuals that appear at the surface to be unrelated (Horn, Birdwell, & Leedy, 1997). Xu and Chen (2004) compared shortest path algorithms and the breadth-first search algorithm in terms of their abilities to find the strongest association paths in criminal networks. The results show that the paths identified by the shortest path algorithms are more useful for generating investigative leads than those identified by the breadth-first search algorithm.

Given a graph consisting of nodes and links, shortest path algorithms can find optimal paths between any pair of nodes in the graph. The Dijkstra algorithm is the classical method for computing the shortest paths from a single source node to every other node in a weighted graph (Dijkstra, 1959; Helgason et al., 1993). Most other algorithms for solving shortest path problems are based on the Dijkstra algorithm but have improved data structures for implementation (Evans & Minieka, 1992).

Association Visualization

Most existing link analysis tools provide an association visualization function. The first link analysis tool is the Anacapa charting system designed to help investigators analyze relationships among a set of directly related people such as members from a gang (Harper & Harris, 1975). With this approach, an investigator first reads various documents to identify relationships between crime entities under study. The results are then assembled into an association matrix in which rows or columns represent individuals. The investigators can draw a link chart based on the association matrix in order to discover new investigative directions or confirm initial suspicions about specific suspects (Sparrow, 1991). Since the Anacapa charting approach was introduced, it has been used widely in all levels of law enforcement and intelligence agencies and has been shown to be very useful in crime investigation. However, this approach is manual in nature and depends on human investigators to identify and chart crime associations. It offers little help in addressing the information overload problem facing link analysis.

Recent years have seen the emergence of many commercial software packages for link visualization such as Analyst's Notebook, Netmap, Crime Link, Orion, and VisuaLink

TABLE 1. Features of existing link analysis tools and systems.

Link analysis systems	Association identification	Knowledge engineering	Association path search	Association visualization
COPLINK Detect	Yes	No	Single level	No
Watson	Yes	No	Single level	Yes
FinCEN	Yes	Yes	Single level	Yes
Link Discovery Tool	No	No	Multiple level, shortest path	Yes
Analyst's Notebook, Netmap, etc.	No	No	Single level	Yes

(Mena, 2003). Most of these software packages provide automatic charting and graph layout features to facilitate the visualization of associations between crime entities. Analyst's Notebook, for example, allows a user to create different icons for different types of entities and draw lines to connect related entities. Netmap and VisuaLink can automatically generate a link chart based on input association data. Specifically, Netmap arranges all entities on the peripheral of a circle and uses different colors to represent different types of entities. Straight lines are used to represent the associations between entities. VisuaLink can present entities and their associations in various forms such as circular layout, column layout, and network layout.

These tools have advanced association visualization functionality and can help investigators or analysts to organize and present their analysis results. Some of the tools can even incorporate audio or video files in link charts. However, all these software packages require that entity association information be entered into the system or contained in input data files. They do not address the challenges of association identification, search, and analysis caused by information overload and high branching factors. Table 1 summarizes the functionality of current link analysis tools and systems.

In summary, prior work in link analysis has proposed some approaches to addressing the challenges. However, link analysis remains a difficult problem for crime investigators. In the next section we propose several link analysis techniques and present the design of our automated link analysis system, CrimeLink Explorer, to address some of these problems.

System Design

In our design, we use both co-occurrence analysis and a heuristic approach to identify crime entity associations and to determine the association importance. The co-occurrence analysis estimates the likelihood of a strong association between two crime entities based on the co-occurrence weight. The heuristic approach, on the other hand, uses domain knowledge collected from experts to make inferences about the conditional probability of a strong association. To facilitate association path search, we use Dijkstra's shortest path algorithm to find paths that may consist of hidden, intermediate entities and that are most likely to provide the relationship chain between known entities to uncover investigative leads.

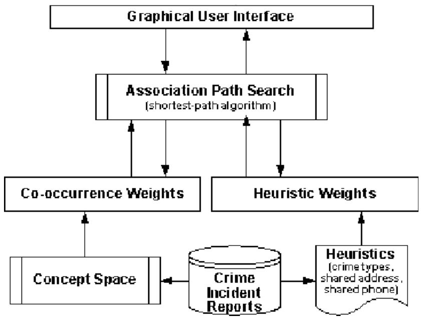


FIG. 1. System architecture.

We designed and implemented a system called CrimeLink Explorer based on a set of structured crime incident data extracted from the Tucson Police Department (TPD) Records Management System. The current version of the system focuses only on associations between persons and does not include other entity types such as location or vehicle. A graphical user interface was developed to allow users to search and visualize association paths. Figure 1 illustrates the system architecture.

Crime Incident Reports

Structured crime incident reports stored in law enforcement databases are a major source of information about both criminal and non-criminal incidents over extended time periods. Incident reports may document serious crimes, such as homicides and robberies, or trivial incidents such as suspicious activity calls or neighbor disputes. Although not as damaging as serious crimes, trivial incidents may provide important information about associations that can later be used to solve serious crimes.

Each crime incident report contains basic information about a case such as date, time, location, persons, role of each person (e.g., victim, suspect, arrestee, and witness), and properties (e.g., weapons and stolen items). Each crime incident is classified as a specific type (e.g., homicide, aggravated assault, robbery, fraud, auto theft, sexual assault, among others) based on the Uniform Crime Reporting (UCR) code that has been the national standard for case classification and crime reporting since 1930. A four-digit UCR code can include more detailed information than a general crime type description. For example, UCR codes 0401, 0402, 0403, and 0413 respectively specify "Aggravated

Assault Officer–serious", "Aggravated Assault Officer–not serious", "Aggravated Assault Other", and "Aggravated Assault Other–Domestic Violence" under the "Aggravated Assault" type. The UCR reporting standard is gradually being replaced by the National Incident-Based Reporting System (NIBRS; Federal Bureau of Investigation 1992). NIBRS captures specific information about the nature of relationships between persons involved in an incident, which may significantly reduce the difficulty of association identification. However, the majority of law enforcement agencies, including TPD, are still using UCR for crime reporting, thus the data in our research does not have explicit information about criminal associations. In addition to incident reports, our data set also contains information about all persons that have been involved in at least one crime incident.

These incident report records and person information are the data source for automated link analysis in this research. Because no explicit information about criminal associations is available in the structured incident data and person data, we use two different approaches, the co-occurrence analysis approach and the heuristic approach, to identify and estimate the importance of criminal associations.

Co-Occurrence Analysis

Originally designed for automatic thesaurus creation, co-occurrence analysis computes a co-occurrence weight between two phrases, i.e., the frequency that the two phrases appear together in the same document (Chen & Lynch, 1992):

$$W_{jk} = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \tag{1}$$

$$W_{kj} = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ik}} \tag{2}$$

The resulting co-occurrence weights are asymmetric, i.e., the co-occurrence weight from phrase j to k, W_{jk}, may not be the same as the co-occurrence weight from phrase k to j, W_{kj}. In equations (1) and (2), d_{ij} indicates whether phrase j appears in document i, d_{ik} indicates whether phrase k appears in document i, and d_{ijk} indicates both phrases j and k are in document i. In this research, we took the average of W_{jk} and W_{kj} for the symmetric co-occurrence weight between phrase j and k.

To apply co-occurrence analysis to criminal association identification, we treated each incident report as a document and each person name as a phrase. We then calculated the co-occurrence weights based on the frequency that two persons appeared together in the same crime incidents and used the co-occurrence weight to represent the strength of an association.

More formally, the criminal association can be modeled as a weighted, undirected graph $G = (V, E)$ with weight function w, where V is the set of nodes and E is the set of edges in the graph. A node $v \in V$ represents a person identified in the incident reports. An edge $(u, v) \in E$ represents a nonzero co-occurrence weight between the nodes u and v. In other words, an edge $(u, v) \in E$ if and only if the two persons u and v co-occurred in at least one incident report. The weight function w: $E \rightarrow \mathbf{R}$ can then be defined based on the co-occurrence weight:

$$w(u, v) = \frac{W_{uv} + W_{vu}}{2} \text{ for all } (u, v) \in E \tag{3}$$

One limitation of the co-occurrence is that a nonzero weight may result by coincidence. An example is a burglary case in which the victim and the suspect appear together but they may have never met. Moreover, in previous studies co-occurrence weights have been found to be of only minor value when subjected to user evaluation. They were shown to be different from investigators' assessments of the strength of criminal associations and were often ignored. Thus, we chose to incorporate heuristics used by crime investigators into our automated link analysis system.

Heuristic Approach

To make a judgment about whether an association is important or strong enough for uncovering investigative leads, crime investigators often use many heuristics based on their experience and knowledge. We interviewed seven crime analysts, two crime intelligence officers, and one police detective sergeant at TPD in order to collect these heuristics for knowledge engineering purposes. The interviewees have been serving in law enforcement for an average of 18 years and they specialize in the investigation of one or more types of crimes, including homicide, aggravated assault, robbery, fraud, auto theft, sexual assault, child sexual abuse, and domestic violence. The total number of crime types considered in our research (based on the four-digit UCR classification code) is 132.

Based on the interviews, we identified three factors which investigators considered while making judgments and the decision rules they used to make judgments.

Sharing addresses or telephone numbers. When two persons have the same home address or phone number, they may be family members, roommates, close friends, or in other close relationships. Therefore, two persons sharing an address or phone number for a period of time can indicate a strong relationship. However, addresses and phone numbers recorded in police databases are often subject to errors such as data entry error and criminal identity deception. As a result, shared addresses and phone numbers are not always reliable. Taking into consideration this data-quality problem, investigators assigned 15% probability to this factor. Representing the strength of an association as the outcome variable O and "two persons have a shared address or phone number" as variable X_1, this probability is formulated as $P(O = \text{strong} \mid X_1 = \text{true}) = 0.15$.

TABLE 2. Probability of strong relationship in a victim-suspect pair for some crime types

UCR	Description	Probability
101	Homicide	97%
102	Manslaughter	80%
201	Rape female	70%
202	Rape male	70%
203	Attempted rape	70%
301	Highway robbery	2%
302	Commercial house robbery	2%
303	Service station robbery	2%
304	Convenience store robbery	2%
305	Residence robbery	5%
306	Bank robbery	1%
307	Misc robbery	2%

Person role combined with crime type. Persons involved in a crime may take different roles: Suspect, Arrestee, Victim, Witness, and Other. All of the crime investigators agreed that most co-arrestees or suspects in an incident had a strong association. Other role pairs, however, varied considerably depending on the type of crime. We collected investigators' probability estimate of each role pair and crime type combination based on investigators' estimation of the strength of the association occurring for that role pair and crime type out of every 100 incidents. For instance, the homicide detective sergeant estimated that about 97 out of 100 homicide incidents included a victim and a suspect who knew each other well or were at least acquaintances. Thus, the corresponding probability for victim-suspect pair in homicide crimes was set to be 97%. Representing the role pair and crime type combination variable as X_2, this means $P(O = \text{strong} \mid X_2 = \text{victim-suspect in homicide}) = 0.97$.

We constructed a probability distribution table for all the probabilities conditional on role pair and crime type combination. The resulting distribution table covers all major role pairs, except for arrestee-arrestee and suspect-suspect pairs, across the 132 crime types we selected. (The two excluded role pairs were recorded in a separate table because their conditional probability for a strong association was assigned 0.99 regardless of crime type.)

Part of the probability table is shown in Table 2. The table shows the UCR codes, the description of crime types, and the associated probabilities of a strong relationship in a victim-suspect role pair in different types of crimes. As can be seen, even for the same role pair the probability of a strong relationship between two persons still varies considerably. For example, for crime types such as homicide, manslaughter, and rape, it is often likely that the suspect and the victim know each other or are acquaintances. Therefore, investigators assigned a high probability to such types of crimes. On the other hand, in crime types such as highway robbery or other types of robbery, the suspects often choose victims at random and it is quite unlikely that the suspect and the victim have a strong relationship.

Repeated co-occurrences in crime incidents. When two persons appear together in multiple incidents, the likelihood of a strong relationship between the two persons is high even if other information does not indicate the existence of a strong relationship. This is the same as the rationale behind the co-occurrence analysis approach. However, because previous user studies have shown that crime investigators do not agree on co-occurrence weights, we estimated the strength of an association resulting from multiple co-occurrences in incidents based on an empirically derived probability distribution.

We obtained the empirical distribution by analyzing a random sample of 40 incident reports of various crime types and counting the number of times each pair of persons co-occurred. We read supporting narrative reports for each incident to determine whether an association was actually a strong relationship such as kinship, close friendship, co-workers, etc. We found that the more times two persons appeared together, the higher the likelihood that they were involved in family related crimes, i.e., the two persons were family members. For example, in 21 out of 40 incidents containing persons who appeared together four times, 15 were domestic violence incidents, custodial interference, or family fights, and the other six were court-ordered enforcements or civil matters that were often related to domestic situations.

Based on our analysis, we constructed the probability distribution by assigning 1% to a single co-occurrence, indicating that it could be completely random with no other facts to support a stronger association. From two to three co-occurrences the probability increased rapidly. The probability distribution above four exceeded 99%, so all pairs of subjects who co-occur four or more times were given a probability of 100%. Figure 2 shows this empirically derived probability distribution of $P(O = \text{strong} \mid X_3 = x)$, where X_3 is the co-occurrence count which may take on any positive integer value.

These three factors were considered the most important by all investigators we interviewed. Some investigators also mentioned other factors, such as shared vehicles and weapons, which they might consider during judgment. To limit the scope of our research we did not consider those factors.

Based on graph theory, we can also represent the criminal network using a weighted, undirected graph $G = (V, E)$ with

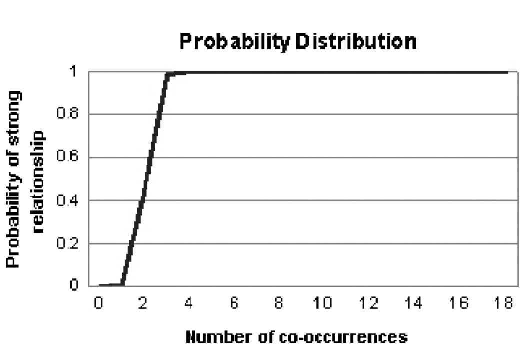


FIG. 2. Association probability distribution vs. the number of co-occurrences.

weight function w, where V is the set of nodes and E is the set of edges in the graph. A node $v \in V$ represents a person identified in the reports and an edge $(u, v) \in E$ represents a nonzero heuristic weight between the nodes u and v. The weight function $w: E \to \mathbf{R}$ can then be defined based on the probabilities discussed above:

$$w(u, v) = P(O = \text{strong} \mid u, v) \text{ for all } (u, v) \in E \quad (4)$$

Association Path Search

We used Dijkstra's shortest path algorithm (Dijkstra, 1959) to find the strongest association paths between two or more known criminals. The original Dijkstra algorithm finds the shortest paths from a single source node to all the other nodes in a graph. It works by maintaining a shortest path tree T rooted at a source node, say s. T contains nodes whose shortest distance from s is already known. Initially, T contains only s. At each step, we select from the candidate set a node with the minimum distance to s and add this node to T. Once T includes all nodes in the graph, the shortest paths from the source node s to all the other nodes have been found.

To apply the shortest path algorithm for finding the strongest paths, we had to address two representation problems. First, in contrast with the traditional shortest path algorithm, a high weight (strong association) is preferred to a low weight in our criminal association network. Second, the strength of an association path between two nodes should be calculated as the product of the weights of all edges in the path rather than the sum of the weights. The reason for this is that a link weight should be treated as a probability measure, which indicates how likely it is that two nodes are related. In general, the probability of a set of mutually independent events occurring together is the product of the probabilities of the individual events. If two nodes are connected by a path consisting of a sequence of edges, the strength of the association between these two nodes should be the product of the weights of these intermediate links.

To address these problems, we used the logarithmic transformation on the path weights (co-occurrence weights based on equation (3) or heuristic weights based on equation (4)) to represent the "distance" between each node (Xu & Chen, 2004). The transformed weight was calculated as follows:

$$d(u, v) = -\ln(w(u, v)) \quad (5)$$

where $w(u, v)$ is the association weight and $d(u, v)$ is the transformed weight for an edge (u, v).

Given this transformation, we postulate the following axioms:

> Axiom 1: All link weights in the new graph are nonnegative numbers.
>
> Axiom 2: A lower link weight in the new graph corresponds with a higher link weight in the original network.
>
> Axiom 3: The shortest path (using summation of link weights) between a pair of nodes in the new graph generates a path with the maximum link weight product among all the alternative paths between these two nodes in the original network.

The proofs of these axioms are fairly straightforward. Readers are referred to (Xu & Chen, 2004) for the complete proofs. Axiom 1 guarantees that the transformed graph does not contain negative-weight links, which is a necessary condition for the shortest path algorithms (Evans & Minieka, 1992). Axioms (2) and (3), respectively, address the two representation problems discussed earlier. Therefore, with such a transformation, we are able to use conventional shortest path algorithms to identify the strongest associations between a pair of nodes or entities in a criminal network.

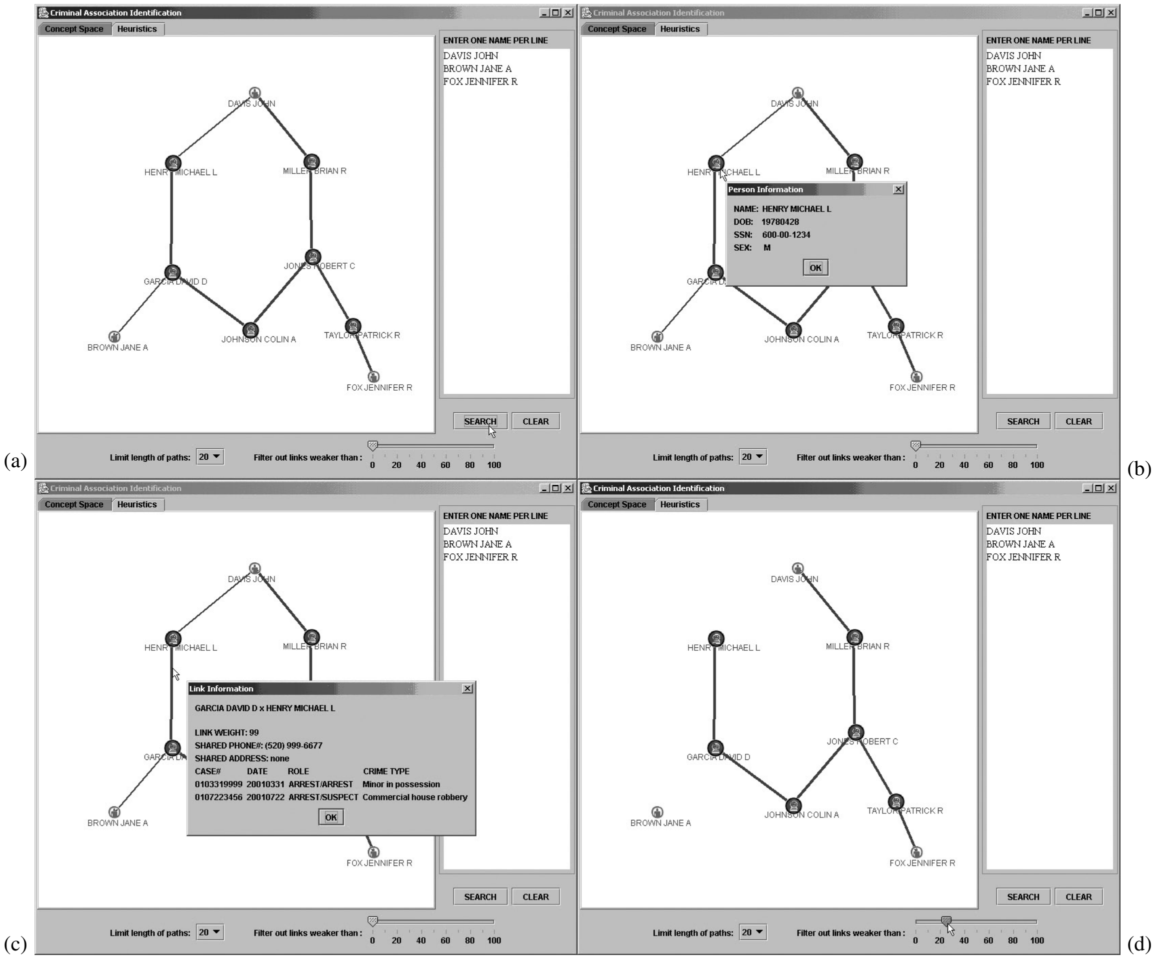
Following the transformation, we can formulate our problem as follows: Given a graph $G = (V, E)$ that represents a criminal network, find the strongest association paths between two or more criminals in the network. As discussed above, the overall weight of a strong path should be the product of the weights of all intermediate links on the path. In other words, given two nodes s and t, we would like to find an acyclic subset $S = \{(s, v_1), (v_1, v_2), \ldots (v_i, t)\} \subseteq E$ that connects s and t such that the association $\Pi_{(u, v) \in S}\,(d(u, v))$ is maximized.

With minor modifications, we used Dijkstra's algorithm to compute the shortest paths from a single source node to a set of specified nodes (instead of all nodes) in the graph. This reduced the processing time and made the algorithm more efficient (Xu & Chen, 2004).

User Interface

A graphical user interface was implemented to allow a user to interact with the system. Figure 3 (a) shows the user interface after the user conducted a search for association paths among three persons.[1] The user entered the names of interest in the text field at the right-hand side and then pressed the search button. The system conducted the shortest path search based on either the co-occurrence weights or heuristic weights depending on the user's choice. The user could then visualize the association paths on the link chart at the left-hand side. Each node on this chart represents a person labeled by the person's name. A straight line between two nodes indicates that the two persons are associated. The thickness of the line is proportional to the association weight. Considering the fact that alternative paths other than the shortest may also exist between two nodes, the system allows user to expand a node (by right-clicking on it) in order to manually explore and construct other paths.

[1] All names and information used in the example have been scrubbed for confidentiality.



ENTER ONE NAME PER LINE
DAVIS JOHN
BROWN JANE A
FOX JENNIFER R

Limit length of paths: 20

Filter out links weaker than :

Person Information

NAME: HENRY MICHAEL L
DOB: 19780428
SSN: 600-00-1234
SEX: M

OK

Link Information

GARCIA DAVID D x HENRY MICHAEL L

LINK WEIGHT: 99
SHARED PHONE#: (520) 999-6677
SHARED ADDRESS: none

```
CASE#       DATE      ROLE             CRIME TYPE
0103319999  20010331  ARREST/ARREST    Minor in possession
0107223456  20010722  ARREST/SUSPECT   Commercial house robbery
```

OK

FIG. 3. Graphical user interface of the CrimeLink Explorer system: (a) the user performs a search on three persons; (b) the user looks for additional information about "Henry Michael L"; (c) the user investigates the association between "Henry Michael L" and "Garcia David D"; (d) the user filters out low-weight associations in the display.

To see additional information (e.g., sex, date of birth, and social security number) about a person, the user can click on the corresponding node. In our example, suppose the user was interested in the person Henry Michael L. The user could then click on this person's name and see his personal details, such as date of birth, social security number, and gender (see Figure 3(b)). The user would like to further investigate this person and noticed that there was a strong relationship between him and Garcia David D (as indicated by the thick line between the two persons on the graph). To see the details of their relationship, the user could also double-click on the link between these two persons to see their association information, including the history of all the past incidents the two persons were involved in together, whether they had shared phone numbers or addresses, and the association weight assigned by the system (see Figure 3(c)). In our example, the two persons did share the same phone number but not the same address. The association weight was 99, which was a high weight.

The bottom panel of the interface allowed the user to set a limit on the number of intermediate links on a path. If the shortest path algorithm could not find a path within the chosen limit number of links between the search targets, the system would indicate that there was no significant path found. The user could also filter out unimportant associations from the link chart by adjusting the slider at the bottom panel (see Figure 3(d)). In the example, the user set the filter to 25, so only links with a link weight of 25 or stronger were shown.

Research Questions

As discussed earlier, with the increasing amount of information available to crime investigators, the information overload problem in crime analysis remains an urgent issue to be resolved. Although various link analysis tools exist, finding the indirect associations between persons involved in a crime is still a difficult and time-consuming task. In this research we aimed to study the following research questions:

- Does the proposed automated link analysis approach help address the challenges of link analysis in the law enforcement domain?
- Does the multilevel search based on shortest path algorithm help users identify associations between persons more efficiently?

DOI: 10.1002/asi

- Does the knowledge engineering approach that incorporates heuristics (human knowledge) into the analysis improve the users' performance in crime investigation?

System Evaluation

Attempting to answer the above questions, we conducted a user study at TPD with ten crime investigators participating. In this section we describe the study in detail and discuss our findings.

Experimental Design

Dataset. When extracting the dataset to use in the experiment, we considered two factors: The dataset must contain (a) real data so that crime investigators would be engaged and interested in the results and (b) sufficient amounts of data for association paths between a reasonable number of subjects to exist. Based on these considerations, we extracted approximately 20 months of incident reports from the TPD database. The resulting dataset contained 239,780 incident reports in which 229,938 persons were involved. Information for each person, such as age, gender, race, address, and phone numbers, was also extracted. The total number of associations among these persons was 207,907. The average branching factor was 0.9, and the maximum branching factor was as high as 107.

Hypotheses and performance metrics. To test our system's abilities to address the problems of information overload and association path search complexity, we compared our system and COPLINK Detect (Hauck et al., 2002) in terms of their efficiencies. We chose the COPLINK Detect system based on two considerations. First, COPLINK Detect was a well-developed commercial product that had been shown to be rather effective and efficient in searching for criminal associations (Hauck et al.). It is a representative of single-level link analysis tools, as it could find crime entities that were directly associated with a given entity. Second, most of our subjects (see later description) had experience with COPLINK Detect, which was the primary tool for crime analysts at TPD to conduct link analysis in their daily work. These allowed us to compare the two systems on a fair base. In addition, we compared the two different weight calculation approaches used in our system, namely the heuristic (knowledge engineering) approach and the co-occurrence analysis. In order to allow fair comparison, both COPLINK Detect and our system were connected to the same dataset described in the Dataset section.

We posed the following hypotheses in our experiment:

> *H1: Subjects will achieve higher efficiency conducting an association path search with CrimeLink Explorer than with the "single-level" link analysis tool.*

Because COPLINK Detect did not facilitate the search for association paths between indirectly connected crime entities, crime investigators had to expand links manually to find possible criminal associations. Our system, in contrast, could search for the strongest association paths between crime entities for multiple levels. The efficiency was defined as the time a subject spent in completing a given task.

> *H2: Association paths found based on the heuristic (knowledge engineering) approach will be more accurate than paths found based on co-occurrence analysis.*

In identifying criminal associations and determining their importance weights, our system could use either the heuristic approach, which captured the domain knowledge crime investigators relied on, or the co-occurrence analysis approach. We hypothesized that heuristic weights would more accurately reflect human judgment than co-occurrence weights, and thus result in more accurate association paths. By having the experts evaluate the outcome based on these two approaches to association-weight calculation, we also measured how practical the approaches were in finding the associations that human experts could use in their crime investigation work.

To measure the accuracy of an association path, we asked subjects to indicate on a 7-point Likert scale how much they agreed with the weights of associations on the path. The accuracy of a path was thus defined as the average agreement scale the subjects indicated on the weights of associations on a path.

> *H3: Subjects will perceive the heuristic approach to be more useful than the co-occurrence approach for investigative work.*

Following the line of H2, we hypothesized that users would find the heuristic approach more useful for investigative work because it incorporated human knowledge from domain experts. Such knowledge would be more helpful for generating leads in their investigation work.

Tasks. We recruited from TPD all available crime analysts and crime intelligence officers, who were familiar with or interested in link analysis. The total number of subjects is 13. We could not have more subjects because other officers and personnel were either not responsible for crime analysis or not interested in participating in research projects. The subjects' average time in their current position is eight years and several subjects were very experienced in link analysis in crime investigation. Most subjects had prior experience in using COPLINK Detect; none of these subjects were in the group of experts whom we interviewed to generate the heuristic rules discussed in the Heuristic Approach section. Each subject was asked to perform three tasks during each experiment:

> *Task 1: Given three person names, use COPLINK Detect to find the strongest association paths among these persons.*

> *Task 2: Given three other person names, use CrimeLink Explorer to find the strongest association paths among these persons based on the co-occurrence weights.*

Task 3: Given the same set of person names used in Task 2, use CrimeLink Explorer to find the strongest association paths among these persons based on the association weights generated by the heuristic approach.

Among the 13 subjects, nine of them did the tasks twice and four of them did only once, so there were 22 evaluation data in total. The four subjects, who did the tasks once, could not participate in the second evaluation due to their busy work schedules. Note that the nine subjects, who did the tasks twice, were given completely different name sets during the second evaluation even though the tasks were the same.

Based on our pilot testing with the system, we found that two-name paths would involve only single link search, which is a relatively simple task in CrimeLink Explorer, but it could be a difficult task in COPLINK Detect when there is no short path between the two person names. Also, crime analysis work often involves more than two persons in practice. On the other hand, finding the strongest path for four or more persons could be overwhelming and not practical for evaluation as too much time would be required from the subjects. Therefore, we chose to use a set of three person names in each task. Two different name sets were used for the three tasks. The tasks and name sets were assigned to the subjects in rotation in order to avoid possible training effects relating to the tasks and the data. Each name set resulted in three association paths connecting the three persons of interest and the number of intermediate links on these association paths ranged from two to five. For each name set, the co-occurrence analysis approach and the heuristic approach found the same paths. However, association weights assigned to individual links using the two approaches were quite different. For example, one of the links of interest received a low weight (14%) from the co-occurrence analysis because the corresponding two persons co-occurred only once. However, the heuristic weight was 99% because the two persons were both arrestees in a commercial house robbery case. Among the nine links on the three paths of that name set, seven links received higher weights from the heuristic approach than from the co-occurrence analysis approach.

In Tasks 2 and 3, each subject was asked to indicate for each link on an association path how much he or she agreed with the weight generated by the system. A 7-point Likert scale was used with one representing strong disagreement and seven representing strong agreement. The agreement scales on individual links were averaged for each path.

The time a subject spent on each task was recorded. The longest time a subject was allowed to complete a task was 30 minutes. Subjects were allowed to give up if they felt completely overwhelmed and were unable to continue a task.

Subjects were asked to complete a post-test questionnaire after they finished all three tasks. Questionnaire items were intended to assess subjects' perception and attitudes toward the usefulness of the system and to collect other comments.

TABLE 3. Experiment results: (a) comparing CrimeLink Explorer and COPLINK Detect; (b) comparing the heuristic approach and the co-occurrence analysis approach.

(a)

System	CrimeLink Explorer	COPLINK Detect	*t*-test *p*-value
Average time spent (sec)	36.5	1416.0	<0.001

(b)

Approach	Heuristic	Concept Space	*t*-test *p*-value
Accuracy (7 is the best)	4.83	2.86	<0.001
Usefulness (7 is the best)	5.63	3.84	0.007

Results and Discussion

We collected the data from the experiment and performed two-tailed *t*-tests to compare the data. The results are summarized in Table 3. In summary, all three hypotheses were supported by the *t*-tests.

H1: Efficiency

H1 was supported ($t = 15.33$, $p < 0.001$). The subjects spent an average of 1416.0 seconds finding the paths among the three persons with COPLINK Detect, while they only spent 36.5 seconds when using CrimeLink Explorer (see Table 3(a)). Most subjects were able to find direct associations of the three given person names using COPLINK Detect, but could not keep track of all the associations that could possibly be generated as they traversed into the second and third level of the search. In 10 of the 22 evaluation tests for Task 1 the subjects were not able to complete the task using COPLINK Detect within 30 minutes. The subjects in these cases indicated that they would need much more time to find and chart association paths between given person names. In contrast, all subjects could quickly find association paths for Task 2 using CrimeLink Explorer. Some of the searches were completed in as few as two seconds; none of the searches took more than 60 seconds.

This result shows that the branching factor increased the search complexity of Task 1 dramatically, making it difficult for subjects to keep track of all possible associations while manually expanding a path. However, the association path search functionality in our system automated the search process using the shortest path algorithm and significantly increased the efficiency.

H2: Accuracy

H2 was supported ($t = 10.38$, $p < 0.001$). Table 3(b) shows that, on average, the subjects agreed with the weights calculated by the heuristic approach in Tasks 2 and 3 with a score of 4.83 out of 7, while the score is only 2.86 for the co-occurrence analysis approach. Figure 4 shows that most

DOI: 10.1002/asi

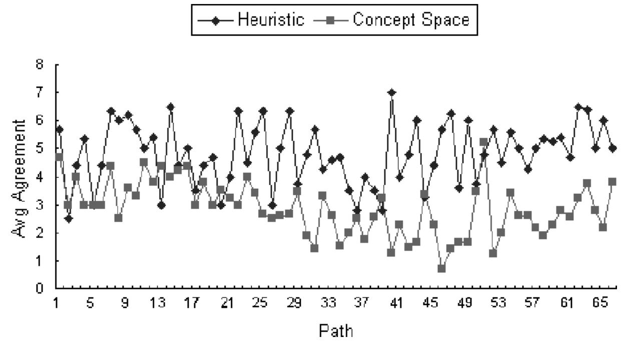


FIG. 4. Agreement differences between the co-occurrence analysis approach and the heuristic approach.

subjects consistently agreed on heuristic weights more often than co-occurrence weights. This means that heuristic weights reflect human judgment more accurately than simple co-occurrence weights, because the heuristic approach incorporates the domain knowledge of crime investigators.

H3: Usefulness

H3 also was supported ($t = 3.06$, $p < 0.01$). Subjects were asked to indicate how much they agreed that the heuristic approach or the co-occurrence analysis approach would be useful in generating investigative leads. The average agreement scales on the heuristic approach and the co-occurrence analysis approach were 5.63 and 3.84, respectively (see Table 3(b)). There were only five cases (out of 66 paths) in which the subjects assigned higher agreement scales to the co-occurrence analysis approach than the heuristic approach. One user who chose the co-occurrence analysis approach over the heuristic approach was later found to have misunderstood the question.

User Feedback and Comments

In addition to the quantitative results, we collected user feedback and comments during the experiment. In general, the subjects were satisfied with CrimeLink Explorer and provided many comments about the strength of the tool. The results are summarized as follows:

Reducing association task complexity. Most subjects believed that the computer automated path search was the only way they could accomplish the task in a reasonable amount of time. They indicated that searching for association paths between crime entities that were not directly connected was too difficult using single-level link analysis. When performing Task 1 using COPLINK Detect, many subjects became frustrated and made comments such as “There was no way I could keep track of all of it”, or “There were too many names. I got lost.” They said it would take them hours or possibly more than a day to find the paths between the persons. One analyst said she had been recently asked to find an association path between two persons. She spent many hours searching and charting links but was never able to find a significant association path because of the large number of possible associations.

Matching expert judgment. Subjects liked the heuristic approach more than the co-occurrence analysis approach. By considering the link information (incident type, person roles, shared phone number, and shared address) provided by the system, the subjects could evaluate how much the link weights matched their own judgment. Some subjects commented on the factor of person role combined with crime type, “That makes more sense, since it takes into account the kind of case.”

Automated system support. All subjects expressed enthusiasm about our system. They believed that such a tool could save them a lot of time on link analysis and uncover important investigative leads. The visualization of association paths would also be very helpful for showing criminal relationships in court. Several subjects asked when they would be able to use the system for their daily work.

Limitations

The current research has several limitations. First, only sets of three names were used in the experiment. The proposed methods being evaluated may perform differently if the number of names in the sets is varied, such as two-name or four-name sets. It would also be interesting to study how the systems perform at different levels of associations between the names. Another limitation is that only person names were used in this evaluation. Other entities, such as addresses and vehicles, could possibly improve performance of the systems. Lastly, the systems were only tested on the data from the Tucson Police Department. Caution needs to be used when applying the model to other datasets.

Conclusions and Future Work

Link analysis faces challenges, such as information overload, association path search complexity, and reliance on domain knowledge. Several techniques were proposed in this article for automated link analysis, including co-occurrence analysis, the shortest path algorithm, and a heuristic approach that captured domain knowledge for determining importance of associations. We implemented the proposed techniques in the prototype CrimeLink Explorer system.

Our system evaluation focused on the system’s efficiency, accuracy, and usefulness, all of which are desirable features of a sophisticated link analysis system. The results of our system evaluation were quite encouraging. The automated link analysis approaches applied in the research could help address the challenges of link analysis and increase the efficiency and accuracy of association path search. Specifically, the heuristic approach was preferred to the co-occurrence analysis approach, because it reflected human judgment more accurately and was more useful for uncovering investigative leads. Moreover, the shortest path algorithm greatly reduced the time crime investigators spent in association path search. Although we only use existing techniques, this research demonstrates that a combination of them can result in increased job productivity and faster crime resolution.

One of the major contributions of the current research is the application of the shortest path algorithm, which has been used mainly in engineering problems such as circuit design, scheduling, and traffic routing, to identify important relations between people in criminal networks. We believe this has opened up some possibilities for future research in applying the algorithm in other areas of social network analysis. In addition, we proposed in this article an alternative way to incorporate human knowledge and decision rules (heuristics) into information systems. In prior knowledge engineering approaches used in information systems such as expert systems, decision rules were coded into the system and the system made a decision directly based on the decision rules. However, in CrimeLink Explorer, heuristics are encoded as relational strength between people, which is a new knowledge engineering approach. In terms of theoretical contribution, we have discussed the application and adaptation of the graph theoretical approach, the shortest path algorithm, in identifying the strongest association path in a criminal network. The model proposed also can be applied to other areas of social network analysis research, such as Internet social network study or business partner analysis.

There are several aspects of the current project that can be improved. For example, we can enhance our system such that it can handle persons whose names were not recorded in the database (e.g., for suspects with only some other attributes available). This will allow the system to find more "hidden" paths between persons that are not identifiable in the current system. It would also be interesting to investigate whether the proposed system performs well under different conditions, e.g., when different numbers of persons are involved or when asymmetric weights are used rather than the averages.

We are currently extending the heuristics to include common vehicles and common organization associates. Such heuristics would be very useful for drawing associations in crime investigations, because some criminals may drive the same vehicle or belong to the same gang. In addition, we plan to analyze the NIBRS (National Incident-Based Reporting System) data (Federal Bureau of Investigation, 1992), which captures specific information about the nature of associations between individuals involved in an incident, to validate the probability tables used in the heuristic approach. Different visualization techniques are also being investigated for improving the system (Xiang, Chau, Atabakhsha, & Chen, 2005). Lastly, we plan to deploy the final version of the CrimeLink Explorer system at the Tucson Police Department such that crime investigators can benefit from the system in solving real-world criminal cases.

Acknowledgments

This project has been primarily funded by the National Science Foundation (NSF), Digital Government Program, "COPLINK Center: Information and Knowledge Management for Law Enforcement," #9983304, July 2000–June 2003 and the NSF Knowledge Discovery and Dissemination (KDD) Initiative. We appreciate the critical and important comments, suggestions, and assistance from Detective Tim Petersen and Dr. Homa Atabakhsh. We also thank all the crime investigators and crime intelligence officers from the Tucson Police Department who participated in our user study.

References

Anderson, T., Arbetter, L., Benawides, A., & Longmore-Etheridge, A. (1994). Security works. Security Management, 38(17), 17–20.

Badiru, A.B., & Cheung, J. (2002). Fuzzy engineering expert systems with neural network applications. New York: Wiley.

Baldwin, B., & Bagga, A. (1998). Coreference as the foundations for link analysis over free text databases. In D. Jensen & H. Goldberg (Eds.), Artificial Intelligence and Link Analysis: Papers from the 1998 Fall Symposium (pp. 8–13). Menlo Park, CA: AAAI Press.

Blair, D.C., & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM, 28(3), 289–299.

Brin, S., & Page, L. (1998, April). The anatomy of a large-scale hypertextual Web search engine. Paper presented at the 7th International Conference on the World Wide Web, Brisbane, Australia.

Brown, D.E., & Hagen, S. (2002). Data association methods with applications to law enforcement. Decision Support Systems 34, 369–378.

Chau, M., Xu, J., & Chen, H. (2002, May). Extracting meaningful entities from police narrative reports. Paper presented at the 2002 Annual National Conference on Digital Government Research, Los Angeles, CA.

Chen, H. (2005). Introduction to the special topic issue: Intelligence and security informatics. Journal of the American Society for Information Science and Technology, 56(3), 217–220.

Chen, H., Chung, W., Xu, J., Wang, G., Chau, M., & Qin, Y. (2004). Crime data mining: A general framework and some examples. IEEE Computer, 37(4), 50–56.

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

Das-Veves, F., Fox, E.A., & Yu, X. (2005, October). Connecting topics in document collections with stepping stones and pathways. Paper presented at the 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany.

Dijkstra, E. (1959). A note on two problems in connection with graphs. Numerische Mathematik, 1, 269–271.

Evans, J., & Minieka, E. (1992). Optimization algorithms for networks and graphs (2nd ed.). New York: Marcel Dekker.

Federal Bureau of Investigation (1992). Uniform crime reporting handbook: National incident-based reporting system (NIBRS). Retrieved January 23, 2007, from http://www.fbi.gov/ucr/nibrs/manuals/v2all.pdf

Fox, M.S., & Smith, S.F. (1984). ISIS: A knowledge-based system for factory scheduling. Expert Systems, 1(1), 25–49.

Goldberg, H.G., & Senator, T.E. (1998). Restructuring databases for knowledge discovery by consolidation and link formation. In D. Jensen & H. Goldberg (Eds.), Artificial Intelligence and Link Analysis: Papers from the 1998 Fall Symposium (pp. 47–52). Menlo Park, CA: AAAI Press.

Goldberg, H.G. & Wong, R.W.H. (1998). Restructuring transactional data for link analysis in the FinCen AI system. In D. Jensen & H. Goldberg (Eds.), Artificial Intelligence and Link Analysis: Papers from the 1998 Fall Symposium (pp. 38–46). Menlo Park: AAAI Press.

Gordon, M., Lindsay, R.K., & Fan, W. (2002). Literature-based discovery on the World Wide Web. ACM Transactions on Internet Technology, 2(4), 261–275.

Goyal, S.K., Prerau, D.S., Lemmon, A.V., Gunderson, A.S., & Reinke, R.E. (1985). COMPASS: An expert system for telephone switch maintenance. Expert Systems, 2(3), 112–126.

Harper, W.R., & Harris, D.H. (1975). The application of link analysis to police intelligence. Human Factors, 17(2), 157–164.

DOI: 10.1002/asi

Hauck, R.V., Atabakhsh, H., Ongvasith, P., Gupta, H., & Chen, H. (2002). Using coplink to analyze criminal-justice data. IEEE Computer, 35(3), 30–37.

Heckerman, D. (1995). A tutorial on learning with Bayesian networks. In M. Jordan (Ed.), Learning in graphical models (pp. 301–354). Cambridge, MA: MIT Press.

Helgason, R.V., Kennington, J.L., & Stewart, B.D. (1993). The one-to-one shortest path problem: An empirical analysis with the two-tree Dijkstra algorithm. Computational Optimization and Applications, 1, 47–75.

Horn, R.D., Birdwell, J.D., & Leedy, L.W. (1997). Link discovery tool. Paper presented at the Counterdrug Technology Assessment Center and Counterdrug Technology Assessment Center's ONDCP/CTAC International Symposium, Chicago, IL.

Jensen, D. (1998). Statistical challenges to inductive inference in linked data. In D. Jensen & H. Goldberg (Eds.), Artificial Intelligence and Link Analysis: Papers from the 1998 Fall Symposium (pp. 59–62). Menlo Park, CA: AAAI Press.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5), 604–632.

Langley, P., & Sage, S. (1994, July). Induction of selective Bayesian classifiers. Paper presented at the 10th Conference of Uncertainty Artificial Intelligence, Seattle, WA.

Lee, R. (1998). Automatic information extraction from documents: A tool for intelligence and law enforcement analysts. In D. Jensen & H. Goldberg (Eds.), Artificial Intelligence and Link Analysis: Papers from the 1998 Fall Symposium (pp. 63–65), Menlo Park, CA: AAAI Press.

Li, K.W., & Yang, C.C. (2005). Automatic crosslingual thesaurus generated from the Hong Kong SAR Police Department Web corpus for crime analysis. Journal of the American Society for Information Science and Technology, 56(3), 272–282.

Lin, S., & Brown, D.E. (2003, June). Criminal incident data association using the OLAP technology. Paper presented at the 1st NSF/NIJ Symposium on Intelligence and Security Informatics (ISI 03), Tucson, AZ.

Lindsay, R.K., & Gordon, M.D. (1999). Literature-based discovery by lexical statistics. Journal of the American Society for Information Science, 50(7), 574–587.

Martin, J., & Oxman, S. (1988). Building expert systems: A tutorial. Englewood Cliffs, NJ: Prentice-Hall.

McAndrew, D. (1999). The structural analysis of criminal networks. In D. Canter & L. Alison (Eds.), The social psychology of crime: Groups, teams, and networks, offender profiling series, III (pp. 53–94). Aldershot, England: Ashgate.

Mena, J. (2003). Investigative data mining for security and criminal detection. Amsterdam, Holland: Butterworth Heinemann.

Quinlan, J.R. (1986). Introduction of decision trees. Machine Learning, 1, 86–106.

Quinlan, J.R. (1993). C4.5: Programs for machine learning. Amsterdam, Holland: Morgan Kaufmann.

Sarkar, S., & Sriram, R.S. (2001). Bayesian models for early warning of bank failures. Management Science, 47(11), 1457–1475.

Shortliffe, E.H. (1976). Computer-based medical consultations: MYCIN. New York: Elsevier, North-Holland.

Sparrow, M.K. (1991). The application of network analysis to criminal intelligence: An assessment of the prospects. Social Networks, 13, 251–274.

Strickland, L.S., & Hunt, L.E. (2005). Technology, security, and individual privacy: New tools, new threats, and new public perceptions. Journal of the American Society for Information Science and Technology, 56(3), 221–234.

Swanson, D.R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30, 7–18.

Swanson, D.R., & Smalheiser, N.R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artificial Intelligence, 91, 183–203.

van der Eijk, C.C., van Mulligen, E.M., Kors, D.A., Mons, B., & van den Berg, J. (2004). Constructing an associative concept space for literature-based discovery. Journal of the American Society for Information Science and Technology, 55(5), 436–444.

Wildemuth, B.M. (2004). The effects of domain knowledge on search tactic formulation. Journal of the American Society for Information Science and Technology, 55(3), 246–258.

Wilson, R., & Sharda, R. (1994). Bankruptcy prediction using neural networks. Decision Support Systems, 11(5), 545–557.

Wu, T., & Pottenger, W.M. (2005). A semi-supervised active learning algorithm for information extraction from textual data. Journal of the American Society for Information Science and Technology, 56(3), 258–271.

Xiang, Y., M. Chau, M., Atabakhsha, H., & Chen, H. (2005). Visualizing criminal relationships: Comparison of a hyperbolic tree and a hierarchical list. Decision Support Systems, 41, 69–83.

Xu, J.J., & Chen, H. (2004). Fighting organized crime: Using shortest path algorithms to identify associations in criminal networks. Decision Support Systems, 38(3), 473–487.

DOI: 10.1002/asi