CHAPTER 6

Intelligence and Security Informatics

Hsinchun Chen and Jennifer Xu
University of Arizona

ISI: Challenges and Research Framework

The tragic events of September 11, 2001, and the subsequent anthrax scare had profound effects on many aspects of society. Terrorism has become the most significant threat to domestic security because of its potential to bring massive damage to the nation's infrastructure and economy. In response to this challenge, federal authorities are actively implementing comprehensive strategies and measures to achieve the three objectives identified in the "National Strategy for Homeland Security" report (U.S. Office of Homeland Security, 2002): (1) preventing future terrorist attacks, (2) reducing the nation's vulnerability, and (3) minimizing the damage and expediting recovery from attacks that occur. State and local law enforcement agencies, likewise, have become more vigilant about criminal activities that can threaten public safety and national security.

Academics in the natural sciences, computational science, information science, social sciences, engineering, medicine, and many other fields have also been called upon to help enhance the government's capabilities to fight terrorism and other crime. Science and technology have been identified in the "National Strategy for Homeland Security" report as the keys to winning the new counter-terrorism war (U. S. Office of Homeland Security, 2002). In particular, it is believed that information technology and information management will play indispensable roles in making the nation safer (Cronin, 2005; Davies, 2002; National Research Council, 2002) by supporting intelligence and knowledge discovery through collecting, processing, analyzing, and utilizing terrorism- and crime-related data (Badiru, Karasz, & Holloway, 1988; Chen, Miranda, Zeng, Demchak, Schroeder, & Madhusudan, 2003; Chen, Moore, Zeng, & Leavitt, 2004). With access to high-quality intelligence, federal, state, and local authorities can make timely decisions to select effective strategies and tactics and to allocate appropriate resources to detect, prevent, and respond to future attacks.

This chapter addresses issues regarding the development of information technologies in the intelligence and security domain. We propose a research framework with a primary focus on knowledge discovery from

databases (KDD). After a comprehensive literature review of existing technologies used in counter-terrorism and crime-fighting applications, we present a set of case studies to demonstrate how KDD and other technologies can contribute to the critical objectives of national security. We also briefly discuss legal, ethical, and social issues related to the use of information technology for national security.

Information Technology and National Security

Six critical mission areas have been identified where information technology can contribute to accomplishing the three strategic national security objectives identified in the "National Strategy for Homeland Security" report (U.S. Office of Homeland Security, 2002):

- *Intelligence and warning.* Although terrorism depends on surprise to bring damage to targets (U.S. Office of Homeland Security, 2002), terrorist activities are neither random nor impossible to track. Terrorists must plan and prepare before the execution of an attack by selecting a target, recruiting and training operatives, acquiring financial support, and traveling to the country where the target is located (Sageman, 2004). To avoid detection, they may hide their true identities and disguise attack-related activities. Similarly, criminals may use falsified identities during police contacts (Wang, Chen, & Atabakhsh, 2004). Although it is difficult, detecting potential terrorist attacks or crimes is possible with the help of information technology. By analyzing communication and activity patterns among terrorists and their contacts, detecting fake identities, and employing surveillance and monitoring techniques, intelligence and warning systems can provide critical alerts and timely warnings to prevent attacks or crimes from occurring.
- *Border and transportation.* Terrorists enter a targeted country by air, land, or sea. Criminals in narcotics rings travel across borders to purchase, transport, distribute, and sell drugs. Information such as travelers' identities, images, fingerprints, and vehicles used is collected from customs, border, and immigration authorities on a daily basis. Such information can greatly improve the capabilities of counter-terrorism and crime-fighting agencies by creating a "smart border," where information from multiple sources is integrated and analyzed to help detect or locate wanted terrorists or criminals. Information sharing and

integration, collaboration and communication, biometrics, and image and speech recognition will all be greatly needed in creating smart borders.

- *Domestic counter-terrorism.* As terrorists may be involved in local crimes, state and local law enforcement agencies also contribute by investigating and prosecuting crimes. Terrorism is regarded as a type of organized crime in which multiple actors cooperate to carry out offenses. Information technologies that help unearth cooperative relationships among criminals and reveal their patterns of interaction would also be helpful for analyzing terrorism.
- *Protecting critical infrastructure.* Roads, bridges, water supply, and many other physical service systems are critical infrastructures and key national assets that may become the target of terrorist attacks because of their vulnerabilities (U.S. Office of Homeland Security, 2002). Moreover, virtual infrastructures such as the Internet are also vulnerable to intrusions and insider threats (Lee & Stolfo, 1998). In addition to physical devices such as sensors and detectors, advanced technologies are needed to model the normal usage behaviors of such systems so that abnormalities and exceptions can be identified. Preemptive or reactive measures can be selected on the basis of the results to secure these assets against attacks.
- *Defending against catastrophic terrorism.* Terrorist attacks can cause devastating damage to a society through the use of chemical, biological, or radiological weapons. Biological attacks, for example, may cause contamination, outbreaks of infectious disease, and significant loss of life. Information systems that can efficiently and effectively collect, access, analyze, and report data about potentially catastrophic events can help agencies prevent, detect, respond to, and manage such attacks (Damianos, Ponte, Wohlever, Reeder, Day, Wilson, et al., 2002).
- *Emergency preparedness and response.* In case of a national emergency, prompt and effective responses are critical to damage containment and control. In addition to the systems that are designed to defend against catastrophes, information technologies that help formulate, experiment with, and optimize response plans (Lu, Huang, & Shekhar, 2003); train response professionals; and manage consequences are

beneficial in the long run. Moreover, systems that provide social and psychological support to the victims of terrorist attacks can also help society recover from disasters.

Given the importance of information technology to national security, its development for counter-terrorism and crime-fighting applications is of the highest priority, despite the many associated problems and challenges.

Problems and Challenges

Intelligence and security agencies routinely gather large amounts of data from various sources. Processing and analyzing such data, however, have become increasingly difficult. Treating terrorism as a form of organized crime allows us to categorize the challenges into three types:

- *Understanding characteristics of criminals and crimes.* Some crimes may be geographically diffused and temporally dispersed. For instance, transnational narcotics trafficking criminals often live in different countries, states, and cities. Drug distribution and sales occur in different places at different times. This is true of other forms of organized crime (e.g., terrorism, sex trafficking, labor racketeering). As a result, investigations must track and prosecute multiple offenders who commit criminal activities in different places at different times. Given the limited resources at the disposal of intelligence and security agencies, this can be difficult. Moreover, as computer and Internet technologies advance, criminals are committing various types of cybercrime under the guise of ordinary online transactions and communications.
- *Understanding characteristics of crime and intelligence related data.* A significant challenge is the information stovepipe and overload resulting from diverse data sources, multiple data formats, and large data volumes. Unlike other professional disciplines such as marketing, finance, and medicine, in which data can be collected from particular sources (e.g., sales records, companies, patient medical histories), the intelligence and security domain does not have a well-defined set of data sources. Both authoritative information (e.g., crime incident reports, telephone records, financial statements, immigration and custom records) and open source information (e.g., news stories, journal articles,

books, Web pages) need to be gathered for investigative purposes. Data collected from these different sources often exist in different formats, ranging from structured database records to unstructured text, image, audio, and video files. Important information such as evidence of criminal associations may be available but buried in unstructured texts and difficult to access and retrieve. Moreover, as data volumes continue to grow, extracting valuable and credible intelligence and knowledge becomes more difficult.

- *Developing crime and intelligence analysis techniques.* Current research on the technologies for counter-terrorism and crime-fighting applications lacks a consistent framework to address the major challenges. Some information technologies, including data integration, data analysis, text mining, image and video processing, and evidence combination, have been identified as particularly helpful (National Research Council, 2002). However, the question of how to employ them in the intelligence and security domain remains unanswered.

We believe that there is a pressing need to develop a science of "Intelligence and Security Informatics" (ISI) (Chen, Miranda, et al., 2003; Chen, Moore, et al., 2004), with its main objective being the "development of advanced information technologies, systems, algorithms, and databases for national security related applications, through an integrated technological, organizational, and policy-based approach" (Chen, Miranda, et al., 2003, p. v).

In comparing ISI with biomedical informatics, a young discipline addressing information management issues in biological and medical applications, we have found important similarities. In terms of data characteristics, they both face the information stovepipe and information overload problems; in terms of technology development, they both are at the exploratory stage of searching for new approaches, methods, and innovative use of existing techniques; in terms of scientific contributions, they both may add new insights and knowledge to fields such as computer science and decision science. Table 6.1 summarizes the similarities and differences between ISI and biomedical informatics. Most importantly, just as a consistent framework has emerged in biomedical informatics (Shortliffe & Blois, 2000), so ISI needs a framework to guide its research agenda. We believe that the knowledge discovery from databases (KDD) methodology, which has achieved significant success in other domains, including business, engineering, biology, and medicine, could be critical in addressing the challenges and problems facing ISI.

Table 6.1 Analogies between ISI and biomedical informatics

		Biomedical Informatics	ISI
Challenges	Domain-Specific	• Complexity and uncertainty associated with organisms and diseases • Critical decisions regarding patient well-being and biomedical discoveries	• Geographically diffused and temporally dispersed organized crimes • Cybercrimes on the Internet • Critical decisions related to public safety and homeland security
	Data	Information stovepipe and overload • HL7 XML standard • PHIN MS messaging • Patient records, diseases data, medical images	Information stovepipe and overload • Justice XML standard • Criminal incident records • Multilingual intelligence open sources
	Technology	• Ontologies and linguistic parsing • Information integration • Data and text mining • Medical decision-support systems and techniques	• Information integration • Criminal network analysis • Data, text, and Web mining • Identity management and deception detection
Methodology		KDD	KDD
Contributions	Scientific	• Computer and information science, sociology, policy, legal • Clinical medicine and biology	• Computer and information science, sociology, policy, legal • Criminology, terrorism research
	Practical	• Public health • Patient well-being • Biomedical treatment and discovery	• Crime investigation and counter-terrorism • National and homeland security

The ISI Framework

To address the data and technical challenges facing ISI, we present a research framework with a primary focus on KDD technologies. The framework is discussed in the context of types of crime and security implications.

Crime is the commission of an act that is forbidden or the omission of a duty that is commanded by a public law, thus making the offender liable to punishment under that law. The greater the threat that a particular crime poses to public safety, the more likely it is to be viewed as a national security concern. Some crimes, such as traffic violations, theft, and homicide, lie mainly in the jurisdiction of local law enforcement agencies. Other crimes need to be dealt with by both local law enforcement and national security authorities. Identity theft and fraud, for instance, are related to criminal identity management issues at both local and national levels. Criminals may escape arrest by using false identities; drug smugglers may enter the United States by holding counterfeited passports or visas. Organized crime, such as terrorism and narcotics trafficking, often diffuses geographically and temporally, resulting in common security concerns across cities and states. Cybercrime can pose threats to public safety across multiple jurisdictions because of the nature of computer network technology. Table 6.2 summarizes the different types of crimes, sorted by security level (Chen, Chung, Wu, Chau, & Qin, 2004).

Table 6.2 Types of crime and security concerns

Crime Types			
	Type	**Local Law Enforcement Level**	**National Security Level**
Increasing public influence ↓	Traffic Violations	Driving under the influence (DUI), fatal/personal injury/property damage, traffic accident, road rage	
	Sex Crime	Sexual offenses, sexual assaults, child molesting	Transnational child pornography
	Theft	Robbery, burglary, larceny, motor vehicle theft, stolen property	Theft of national secrets or weapon information
	Fraud	Forgery and counterfeiting, fraud, embezzlement, identity deception	Transnational money laundering, identity fraud, transnational financial fraud
	Property crime	Property crime (e.g., arson) on buildings, apartments	Intentional destruction of or damage to national infrastructures and assets
	Organized Crime	Narcotic drug offenses (sales or possession), gang-related offenses,	Transnational drug trafficking, terrorism (bioterrorism, bombing, hijacking, etc.), organized prostitution
	Violent Crime	Criminal homicide, armed robbery, aggravated assault, other assaults	Terrorism
	Cybercrime	Internet fraud (e.g., credit card fraud, advance fee fraud, fraudulent Web sites), theft of confidential information	Network intrusion/hacking, illegal trading, virus spreading, cyberpiracy, cyberpornography, cyberterrorism, theft of confidential information

We believe that KDD techniques can play a central role in improving the counter-terrorism and crime-fighting capabilities of intelligence and security agencies by reducing cognitive and information overload. Knowledge discovery refers to nontrivial extraction of implicit, previously unknown, and potentially useful knowledge from data. Knowledge discovery techniques promise easy, convenient, and practical exploration of very large collections of data for organizations and users, and have been applied in marketing, finance, manufacturing, biology, and many other domains (e.g., predicting consumer behavior, detecting credit card fraud, or clustering genes that have similar biological functions). Knowledge discovery usually consists of multiple stages, including data selection, data preprocessing, data transformation, data mining, and the interpretation and evaluation of patterns (Fayyad, Piatetsk-Shapiro, & Smyth, 1996). Data mining plays a key role in extracting patterns from data. Traditional data mining techniques include association-rule mining, classification and prediction, cluster analysis, and outlier analysis (Han & Kamber, 2001). As natural language processing (NLP) research advances, text mining approaches that automatically extract, summarize, categorize, and translate text documents are also being widely used (Trybula, 1999).

Many of these KDD technologies could be applied in ISI studies (Chen, Miranda, et al., 2003; Chen, Moore, et al., 2004). We categorize existing ISI technologies into six classes: information sharing and collaboration, crime association mining, crime classification and clustering, intelligence text mining, spatial and temporal crime pattern mining, and criminal network mining. These six classes are grounded in traditional knowledge

discovery technologies, but include a few new approaches, such as spatial and temporal crime pattern mining and criminal network analysis, that are more relevant to counter-terrorism and crime investigation. Although information sharing and collaboration are not knowledge discovery per se, they help integrate, warehouse, and prepare data for knowledge discovery and thus are included in the framework.

We present in Figure 6.1 our proposed research framework with the horizontal axis representing crime types and vertical axis the six classes of techniques (Chen, Chung, et al., 2004). The shaded regions on the chart show promising research areas, that is, certain classes of techniques are relevant to solving certain types of crime. Note that more serious crimes may require a more complete set of knowledge discovery techniques. For example, the investigation of terrorism may depend on criminal network analysis technology, which requires the use of other knowledge discovery techniques such as association mining and clustering. An important observation about this framework is that the high-frequency occurrences and strong association patterns of severe and organized crime, such as terrorism and narcotics, present a unique opportunity and potentially high rewards for adopting a knowledge discovery framework.

Caveats for ISI

Before we review the technical foundations and approaches, we want to discuss briefly the legal and ethical caveats regarding crime and intelligence research. The potential negative effects of intelligence gathering

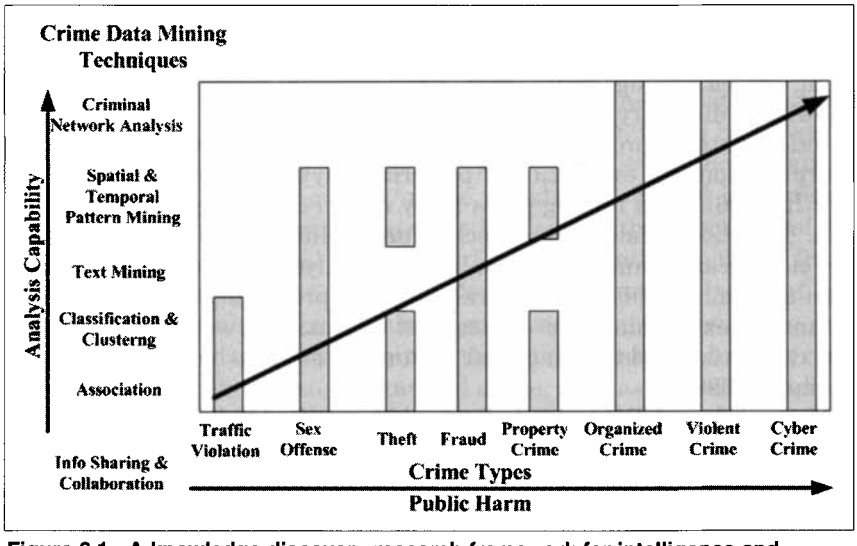

Figure 6.1 A knowledge discovery research framework for intelligence and security informatics.

and analysis on the privacy and civil liberties of the public have been well publicized (Cook & Cook, 2003). Many laws, regulations, and agreements governing data collection, confidentiality, and reporting could influence directly the development and application of ISI technologies. We strongly recommend that intelligence and security agencies and ISI researchers be aware of these laws and regulations in their research efforts (Strickland, 2005). Moreover, we also suggest that a hypothesis-guided, evidence-based approach be used in crime and intelligence analysis research. That is, there should be probable and reasonable causes and evidence for targeting particular individuals or data sets for analysis. Proper investigative and legal procedures need to be strictly followed. It is neither ethical nor legal to "fish" for potential criminals from diverse and mixed crime-, intelligence-, and citizen-related data sources (Strickland, 2005). The well-publicized Defense Advanced Research Program Agency (DARPA) Total Information Awareness (TIA) program and the Multi-State Anti-Terrorism Information Exchange (MATRIX) system, for example, were roundly criticized for their inappropriate use of citizen data and unguided analysis technologies resulting in the potential impairment of Americans' civil liberties (American Civil Liberties Union, 2004). Many new and important privately and publicly funded research projects aim to address these privacy and civil liberties issues in the context of homeland security research. For example, the Electronic Frontier Foundation monitors limits placed on freedom of expression as indicated by Web sites closed for national security reasons by government or Internet service providers (http://www.eff.org/Privacy/Surveillance/Terrorism/antiterrorism_chill.html); the OpenNet Initiative is conducting a three-year study of Internet filtering in Saudi Arabia (http://www.opennetinitiative.net/studies/saudi).

ISI: Technical Foundations and Approaches

In this section, we review the technical foundations of ISI and the six classes of technologies and approaches specified in our research framework. We also summarize relevant past and ongoing research that addresses knowledge discovery in public safety and national security.

Information Sharing and Collaboration

Information sharing across jurisdictional boundaries of intelligence and security agencies has been identified as a key foundation of national security (U.S. Office of Homeland Security, 2002). However, sharing and integrating information from distributed, heterogeneous, and autonomous data sources is a nontrivial task (Hasselbring, 2000; Rahm & Bernstein, 2001). In addition to legal and cultural issues regarding information sharing, it is often difficult to integrate and combine data that are organized in different schemas and stored in different database systems running on different hardware platforms and operating systems

(Hasselbring, 2000). Other data integration problems include: (1) name differences (same entity with different names), (2) mismatched domains (problems with units of measure or reference point), (3) missing data (incomplete data sources or different data available from different sources), and (4) object identification (no global ID values and no inter-database ID tables) (Chen & Rotem, 1998).

Three approaches to data integration have been proposed: *federation*, *warehousing*, and *mediation* (Garcia-Molina, Ullman, & Widom, 2002). Database federation maintains data in their original, independent sources but provides a uniform data access mechanism (Buccella, Cechich, & Brisaboa, 2003; Haas, 2002). Data warehousing is an integrated system in which copies of data from different data sources are migrated and stored to provide uniform access. Data mediation relies on "wrappers" to translate and pass queries from multiple data sources. The wrappers are "transparent" to an application so that the multiple databases appear to be a single database. These techniques are not mutually exclusive and many hybrid approaches have been proposed (Jhingran, Mattos, & Pirahesh, 2002).

All these techniques are dependent, to a great extent, on the matching between different databases. The task of database matching can be broadly divided into *schema-level* and *instance-level matching* (Lim, Srivastava, Prabhakar, & Richardson, 1996; Rahm & Berhstein, 2001). Schema-level matching is performed by aligning semantically corresponding columns between two sources. Various schema elements such as attribute name, description, data type, and constraints may be used to generate a mapping between the two schemas (Rahm & Bernstein, 2001). For example, prior studies have used linguistic matchers to find similar attribute names based on synonyms, common substrings, pronunciation, and Soundex codes (Newcombe, Kennedy, Axford, & James, 1959) to match attributes from different databases (Bell & Sethi, 2001). Instance-level or entity-level matching connects records describing a particular object in one database to records describing the same object in another. Entity-level matching is frequently performed after schema-level matching is completed. Existing entity matching approaches include (1) key equivalence, (2) user-specified equivalence, (3) probabilistic key equivalence, (4) probabilistic attribute equivalence, and (5) heuristic rules (Lim et al., 1996).

Some of these information integration approaches have been used in law enforcement and intelligence agencies for investigation support. The COPLINK Connect system (Chen, Schroeder, Hauck, Ridgeway, Atabakhsh, Gupta, et al., 2003) employed the database federation approach to achieve schema-level data integration. It provided a common COPLINK schema and a one-stop-shop user interface to facilitate access to different data sources from multiple police departments. Evaluation results showed that COPLINK Connect had out-performed the Record Management System (RMS) of police data in system effectiveness, ease of use, and interface design (Chen, Schroeder, et al., 2003).

Similarly, the Phoenix Police Department Reports (PPDR) is a Web-based, federated intelligence system in which databases share a common schema (Dolotov & Strickler, 2003). The bioterrorism surveillance systems developed at the University of South Florida, on the other hand, used data warehouses to integrate historical and real-time surveillance data and incrementally incorporated data from diverse disease sources (Berndt, Bhat, Fisher, Hevner, & Studnicki, 2004; Berndt, Hevner, & Studnicki, 2003). A transnational information-sharing system developed at the University of Florida employed a data mediation approach (Kasad & Su, 2004). The system accessed different databases via a wrapper query processor, which tailored a user query into database-specific queries. This system was intended to enhance information sharing between immigration and border controls in multiple countries.

Integrating data at the entity level has also been difficult. In addition to existing key equivalence matching and heuristic consolidation approaches (Goldberg & Senator, 1998), use of the National Incident-Based Reporting System (NIBRS) (U.S. Federal Bureau of Investigation, 1992), a crime incident classification standard, has been proposed to enhance data sharing among law enforcement agencies (Faggiani & McLaughlin, 1999; Schroeder, Xu, & Chen, 2003). In the Violent Crime Linkage Analysis System (ViCLAS) (Collins, Johnson, Choy, Davidson, & Mackay, 1998), data collection and encoding standards were used to capture more than 100 behavioral characteristics of offenders in serial violent crimes in order to address the problem of entity-level matching.

Information sharing has also been undertaken in intelligence and security agencies through cross-jurisdictional collaborative systems. The COPLINK Agent ran on top of the COPLINK Connect system (Chen, Schroeder, et al., 2003) and linked crime investigators who were working on related cases at different units to enhance collaborations (Zeng, Qin, Huang, & Chen, 2003). It employed collaborative filtering approaches (Goldberg, Nichols, Oki, & Terry, 1992), which have been widely studied in commercial recommender systems, to identify law enforcement users who had similar search histories. Similar search histories might indicate that these users had similar information needs and thus were working on related crime cases. When one user searched for information about a crime or a suspect, the system would alert other users who worked on related cases so that these users could collaborate and share their information through other communication channels. The FALCON system offered similar monitoring and alerting functionality (M. Brown, 1998). Its collaboration capability, however, was relatively limited. The JNET system (http://www.pajnet.state.pa.us/pajnet/site/default.asp) also provides an alerting capability that immediately notifies relevant agencies via pager or e-mail when a wanted person is found or arrested by other agencies. Research has also been performed to model mathematically collaboration processes across law enforcement and intelligence jurisdictions in order to improve work productivity (Raghu, Ramesh, & Whinston, 2003; Zhao, Bi, & Chen, 2003). Although

information sharing and collaboration are not knowledge discovery per se, they prepare data for important subsequent analyses.

Crime Association Mining

Finding associations among data items is an important topic in knowledge discovery research. One of most widely studied approaches is association-rule mining, a process for discovering frequently occurring item sets in a database. Association-rule mining is often used in market basket analysis where the objective is to find which products are bought with which other products (Agrawal, Imielinski, & Swami, 1993; Mannila, Toivonen, & Inkeri, 1994; Silverstein, Brin, & Motwani, 1998). An association is expressed as a rule $X \Rightarrow Y$, indicating that item set X and item set Y occur together in the same transaction (Agrawal et al., 1993). Each rule is evaluated using two probability measures, *support* and *confidence*, where *support* is defined as $prob(X \cap Y)$ and *confidence* as $prob(X \cap Y)/prob(X)$. For example, "diaper $\Rightarrow$ milk with 60 percent *support* and 90 percent *confidence*" means that 60 percent of customers buy both diapers and milk in the same transaction and that 90 percent of the customers who buy diapers tend to buy milk at the same time.

In the intelligence and security domain, spatial association-rule mining (Koperski & Han, 1995) has been proposed to extract cause-effect relations among geographically referenced crime data to identify environmental factors that attract crime (Estivill-Castro & Lee, 2001). Moreover, the research on association mining is not limited to association-rule mining but covers the extraction of a wide variety of relationships among crime data items. Crime association mining techniques can include *incident association mining* and *entity association mining* (Lin & Brown, 2003).

The purpose of incident association mining is to find crimes that might have been committed by the same offender; unsolved crimes are linked to solved crimes to identify the suspect. This technique is often used to solve serial sexual offenses and serial homicides. However, finding associated crime incidents can be time-consuming if it is performed manually. It is estimated that pairwise, manual comparisons on just a few hundred crime incidents would take more than one million human hours (Brown & Hagen, 2002). When the number of crime incidents is large, manual identification of associations between crimes is prohibitively expensive. Two approaches, *similarity-based* and *outlier-based*, have been developed for incident association mining. For example, the Violent Criminal Apprehension Program (ViCAP) identifies similar features or traits of violent crimes such as homicides to detect serial offenders (Icove, 1986).

Similarity-based methods detect associations between crime incidents by comparing features such as spatial locations of the incidents and the offender's modus operandi (MO), often regarded as a criminal's "behavioral signature" (O'Hara & O'Hara, 1980). Expert systems relying

on decision rules acquired from domain experts used to be a common approach to associating crime incidents (Badiru et al., 1988; Bowen, 1994; Brahan, Lam, Chan, & Leung, 1998). However, as the collection of human decision rules requires considerable knowledge engineering effort and the rules collected are often hard to update, the expert system approach has been replaced by more automated approaches. Brown and Hagen (2002) developed a total similarity measure between two crime records as a weighted sum of similarities of various crime features. For features that take on categorical values (such as an offender's eye color), Brown developed a similarity table based on heuristics that specified the similarity level for each pair of categorical values. Evaluation showed that this approach enhanced both accuracy and efficiency for associating crime records. Similarly, Wang, Lin, Shieh, and Deng (2003) proposed measuring similarity between a new crime incident and existing criminal information stored in police databases by representing the new incident as a query and existing criminal information as vector space. The vector space model is widely employed in information retrieval applications; various similarity measures such as the Jaccard function (Rasmussen, 1992) could be used.

Unlike similarity-based methods, which identify associations based on a number of crime features, the outlier-based method focuses only on the distinctive features of a crime (Lin & Brown, 2003). Imagine a series of robberies in which a Japanese sword was used as the weapon. Because a Japanese sword is a very uncommon weapon, unlike, say, a shotgun, it is probable that this series of robberies was committed by the same offender. Based on this outlier concept, crime investigators need first to cluster incidents into cells and then use an outlier score function to measure the distinctiveness of the incidents in a specific cell. If the outlier score of a cell is larger than a threshold value, the incidents contained in the cell are assumed to be associated and committed by the same offender. Evaluation has shown that the outlier-based method is more effective than the similarity-based method proposed in Brown and Hagen (2002).

The task of finding and charting associations between crime entities such as persons, weapons, and organizations is often referred to as entity association mining (Lin & Brown, 2003) or link analysis (Sparrow, 1991) in law enforcement. The purpose is to find out whether crime entities that appear to be unrelated at the surface are actually linked, and if so, how. Law enforcement officers and criminal investigators throughout the world have long used link analysis to search for and analyze relationships among criminals. 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 investigative leads (Schroeder et al., 2003). Although link analysis helps trace criminals through chains of relations, manually identifying and detecting criminal relations from large amounts of criminal-justice data is very time-consuming.

Three types of automated link analysis approaches have been suggested: *heuristic-based*, *statistically-based*, and *template-based*. Heuristic-based approaches rely on decision rules used by domain experts to determine whether two entities in question are related. For example, Goldberg and Senator (1998) suggested that links or associations between individuals in financial transactions be created based on a set of heuristics, such as whether the individuals had shared addresses, shared bank accounts, or related transactions. This technique has been employed by the FinCEN system of the U.S. Department of the Treasury to detect money laundering transactions and activities (Goldberg & Senator, 1998; Goldberg & Wong, 1998). The COPLINK Detect system (Hauck, Atabakhsh, Ongvasith, Gupta, & Chen, 2002) employed a statistically based approach called Concept Space (Chen & Lynch, 1992). This approach measures the weighted co-occurrence associations between records of entities (persons, organizations, vehicles, and locations) stored in crime databases. An association exists between a pair of entities if they appear together in the same criminal incident. The more frequently they occur together, the stronger the association. Zhang, Salerno, and Yu (2003) proposed to use a fuzzy resemblance function to calculate the correlation between two individuals' past financial transactions in order to detect associations between the individuals who might have been involved in a specific money-laundering crime. If the correlation between two individuals is higher than a threshold value, these two individuals are regarded as being related. The template-based approach has been used primarily to identify associations between entities extracted from textual documents, such as police report narratives. Lee (1998) developed a template-based technique using relation-specifying words and phrases. For example, the phrase "member of" indicates an entity–entity association between an individual and an organization. Coady (1985) proposed to use the PROLOG language to derive rules of entity associations automatically from text data and use the rules to detect associations in similar documents. Template-based approaches rely heavily on a fixed set of predefined patterns and rules, and thus may have limited application scope.

Crime Classification and Clustering

Classification is the process of mapping data items into one of several predefined categories based on attribute values of the items (Hand, 1981; Weiss & Kulikowski, 1991). Examples of classification applications include fraud detection (Chan & Stolfo, 1998), computer and network intrusion detection (Lee & Stolfo, 1998), bank failure prediction (Sarkar & Sriram, 2001), and image categorization (Fayyad, Djorgovish, & Weir, 1996). Classification is a type of supervised learning that consists of a training stage and a testing stage. Accordingly, the dataset is divided into a training set and a testing set. The classifier is designed to "learn" from the training set classification models

governing the membership of data items. The accuracy of the classifier is assessed using the testing set.

Discriminant analysis (Eisenbeis & Avery, 1972), Bayesian models (Duda & Hart, 1973; Heckerman, 1995), decision trees (Quinlan, 1986, 1993), artificial neural networks (Rummelhart, Hinton, & Williams, 1986), and support vector machines (SVM) (Vapnik, 1995) are widely used classification techniques. In discriminant analysis the class membership of a data item is modeled as a function of the item's attribute values. Through regression analysis a class membership discriminant function can be obtained and used to classify new data items. Bayesian classifiers assume that all data attributes are conditionally independent, given the class membership outcome. The task is to learn the conditional probabilities among the attributes, given the class membership outcome. The learned model is then used to predict the class membership of new data items based on their attribute values. Decision tree classifiers organize decision rules learned from training data in the form of a tree. Algorithms such as ID3 (Quinlan, 1986, 1993) and C4.5 (Quinlan, 1993) are popular decision tree classifiers. An artificial neural network consists of interconnected nodes to imitate the functioning of neurons and synapses of human brains. It usually contains an input layer with nodes taking on the attribute values of data items and the output layer with nodes representing class membership labels. Neural networks learn and encode knowledge through connection weights. SVM is a novel learning classifier based on the Structural Risk Minimization principle from computational learning theory. SVM is capable of handling millions of inputs and does not require feature selection (Cristianini & Shawe-Taylor, 2000). Each of these classification techniques has its advantages and disadvantages in terms of accuracy, efficiency, and interpretability. Researchers have also proposed hybrid approaches to combine these techniques (Kumar & Olmeda, 1999).

Several of these techniques have been applied in the intelligence and security domain to detect financial fraud and computer network intrusion. For example, in order to identify fraudulent financial transactions, Aleskerov, Freisleben, and Rao (1997) employed neural networks to detect anomalies in customers' credit card transactions based on their transaction histories. Hassibi (2000) employed a feed-forward back-propagation neural network to compute the probability that a given transaction was fraudulent. Two types of intrusion detection, *misuse detection* and *anomaly detection*, have been studied in computer network security applications (Lee & Stolfo, 1998). Misuse detection identifies attacks by matching them onto previously known attack patterns or signatures. Anomaly detection, on the other hand, identifies abnormal user behaviors based on historical data. Lee and Stolfo (1998) employed decision rule induction approaches to classify *sendmail* system call traces into normal or abnormal traces. Ryan, Lin, and Mikkulainen (1998) developed a neural network-based intrusion detection system to detect unusual user activity based on the patterns of users' past system command usage.

Stolfo, Hershkop, Wang, Nimeskern, and Hu (2003) applied Bayesian classifiers to distinguish between normal e-mail and spamming e-mail.

Unlike classification, clustering is a type of unsupervised learning. It groups similar data items into clusters without knowing their class membership. The basic principle is to maximize intra-cluster similarity while minimizing intercluster similarity (Jain, Murty, & Flynn, 1999). Clustering has been used in a variety of applications including image segmentation (Jain & Flynn, 1996), gene clustering (Eisen, Spellman, Brown, & Botstein, 1998), and document categorization (Chen, Houston, Sewell, & Schatz, 1998; Chen, Schuffels, & Orwig, 1996). Various clustering methods have been developed, including *hierarchical approaches,* such as complete-link algorithms (Defays, 1977), *partitional approaches,* such as *k*-means (Anderberg, 1973; Kohonen, 1995), and *Self-Organizing Maps* (SOM) (Kohonen, 1995). These clustering methods group data items based on different criteria and may not generate the same clustering results. Hierarchical clustering groups data items into a series of nested clusters and generates a tree-like dendrogram in which each node represents a merging of clusters. Partitional clustering algorithms generate only one partition level rather than nested clusters. Partitional clustering is more efficient and scalable for large datasets than hierarchical clustering, but has difficulty determining the appropriate number of clusters (Jain et al., 1999). In contrast to the hierarchical and partitional clustering that relies on the similarity or proximity measures between data items, SOM is a neural network-based approach that directly projects multivariate data items onto two-dimensional maps. SOM can be used for clustering and visualizing data items and groups (Chen, Schuffels, et al., 1996).

The use of clustering methods in the law enforcement and security domains can be categorized into two types: *crime incident clustering* and *criminal clustering*. The purpose of crime incident clustering is to find a set of similar crime incidents based on an offender's behavioral traits or a geographical area with a high concentration of certain types of crimes. For example, Adderley and Musgrove (2001) employed the SOM approach to cluster sexual attack crimes based on a number of offender MO attributes (e.g., the precaution methods taken and the verbal themes during the crime) in order to identify serial sexual offenders. The clusters found were used to form offender profiles containing MO and other information such as offender motives and racial preferences when choosing victims. Similarly, Kangas, Terrones, Keppel, and La Moria (2003) employed the SOM method to group crime incidents in order to identify serial murderers and sexual offenders. D. Brown (1998) proposed *k*-means and the nearest neighbor approach to clustering spatial data of crimes to find "hot spot" areas in a city. Spatial clustering methods are often used in "hot spot analysis," which will be reviewed in detail in the section on spatial and temporal mining.

Criminal clustering is often used to identify groups of criminals who are closely related. Instead of using similarity measures, this type of clustering relies on relational strength that measures the intensity and

frequency of relationships between offenders. Stolfo et al. (2003) proposed grouping e-mail users who frequently communicated with each other into clusters so that unusual e-mail behavior that violated the group communication patterns could be detected. Offender clustering is more often used in criminal network analysis, which will be reviewed in detail in the section with that title.

Intelligence Text Mining

A large amount of intelligence- and security-related data is represented in text form such as police narrative reports, court transcripts, news stories, and Web articles. Valuable information in such texts is often difficult to retrieve, access, and use for the purposes of criminal investigation and counter-terrorism. It is desirable to mine the text data automatically in order to discover valuable knowledge about criminal or terrorism activities.

Text mining has attracted increasing attention in recent years as natural language processing capabilities advance (Chen, 2001). An important task of text mining is information extraction, a process of identifying and extracting from free text select types of information such as entities, relationships, and events (Grishman, 2003). The most widely studied information extraction subfield is named entity extraction. It helps to automatically identify from text documents the names of entities of interest, such as persons (e.g., "John Doe"), locations (e.g., "Washington, DC"), and organizations (e.g., "National Science Foundation"). It has also been extended to identify other text patterns, such as dates, times, number expressions, dollar amounts, e-mail addresses, and Web addresses (URLs). The Message Understanding Conference (MUC) series has served as the major forum for researchers in this area to compare the performance of their entity extraction approaches (Chinchor, 1998).

Four major named-entity extraction approaches have been proposed: lexical lookup, rule-based, statistical models, and machine learning.

- *Lexical lookup.* Most research systems maintain hand-crafted lexicons that contain lists of popular names for entities of interest, such as all registered organizational names in the U.S. and all personal surnames obtained from government census data. These systems work by looking up phrases in texts that match the items specified in their lexicons (e.g., Borthwick, Sterling, Agichtein, & Grishman, 1998).
- *Rule-based.* Rule-based systems rely on hand-crafted rules to identify named entities. The rules may be structural, contextual, or lexical (Krupka & Hausman, 1998). An example rule would look like the following:

  capitalized last name + , + *capitalized first name* ⇒ *person name*

Although such human-created rules are usually of high quality, this approach may not be easy to apply to other entity types.

- *Statistical models*. Such systems often use statistical models to identify occurrences of certain cues of particular patterns for entities in texts. A training data set is needed for a system to acquire the statistics. The statistical language model reported in Witten, Bray, Mahoui, and Teahan (1999) is an example of such a system. It uses the Prediction by Partial Matching (PPM) model to extract entities from text based on conditional probability distributions of characters. The probability of occurrence of later characters in a word or phrase depends on the occurrence of preceding characters; for example, "12Jan2005" in a newsletter can be correctly identified as a time phrase using this model.
- *Machine learning*. This type of system relies on machine learning algorithms rather than human-created rules to extract knowledge or identify patterns from textual data. Examples of machine learning algorithms used in entity extraction include neural networks, decision trees (Baluja, Mittal, & Sukthankar, 1999), Hidden Markov Models (Miller, Crystal, Fox, Ramshaw, Schwartz, Stone, et al., 1998), and entropy maximization (Borthwick et al., 1998).

Instead of relying on a single method, most existing information extraction systems combine two or more of these approaches. Many systems were evaluated at the MUC-7 conference. The best systems were able to achieve over 90 percent in both precision and recall rates in extracting persons, locations, organizations, dates, times, currencies, and percentages from a collection of *New York Times* news stories.

Recent years have seen research on named-entity extraction for intelligence and security applications (Patman & Thompson, 2003; Wang, Huang, Teng, & Chien, 2004). For example, Chau, Xu, and Chen (2002) developed a neural network-based entity extraction system to identify personal names, addresses, narcotic drugs, and personal property names from police report narratives. Rather than relying entirely on manual rule generation, this system combines lexical lookup, machine learning, and some hand-crafted rules. The system achieved over 70-percent precision and recall rates for personal names and narcotic drug names. However, it was difficult to achieve satisfactory performance for addresses and personal property because of their wide variation. Sun, Naing, Lim, and Lam (2003) converted the entity extraction problem into a classification problem in order to identify relevant entities from

the MUC text collection on terrorism. They first identified all noun phrases in a document and then used the support vector machine to classify those entity candidates on the basis of both content and context features. The results showed that for the specific terrorism text collection, the performance of this approach in regards to precision and F measure was comparable to AutoSlog (Riloff, 1996), one of the best entity extraction systems reported earlier.

Several news and event extraction systems have been reported recently, such as Columbia's Newsblaster (McKeown, Barzilay, Chen, Elson, Evans, Klavans, et al., 2003) and CMU's (Carnegie Mellon University) system (Yang, Carbonell, Brown, Pierce, Archibald, & Liu, 1999), which automatically extract, categorize, and summarize events from international online news sources. Some of these systems can also work for multilingual documents and have great potential for automatic detection and tracking of terrorism events for intelligence purposes.

Crime Spatial and Temporal Mining

Most crimes, including terrorism, have significant spatial and temporal characteristics (Brantingham & Brantingham, 1981). Analysis of spatial and temporal patterns of crimes continues to be one of the most important crime investigation techniques. It aims to gather intelligence about environmental factors that prevent or encourage crimes (Brantingham & Brantingham, 1981), identify geographic areas of high crime concentration (Levine, 2000), and detect criminal trends (Schumacher & Leitner, 1999). The discovery of such patterns makes possible the use of effective and proactive control strategies, such as allocating the appropriate amount of policing resources in certain areas at certain times, to prevent crimes.

Spatial pattern analysis and geographical profiling play important roles in solving crimes (Rossmo, 1995). Three approaches for crime spatial pattern mining have been reported: *visual approaches*, *clustering approaches*, and *statistical approaches* (Murray, McGuffog, Western, & Mullins, 2001). The visual approach is also called crime mapping. It presents a city or regional map annotated with various crime-related information. For example, a map can be color-coded to present the densities of a specific type of crime in different geographical areas. Such an approach can help users visually detect relationships between spatial features and the occurrence of crime. The clustering approach has been used in hot spot analysis, a process of automatically identifying areas with high crime concentration. This type of analysis helps law enforcement effectively allocate policing resources to reduce crime in hot spot areas. Partitional clustering algorithms such as the k-means methods are often used for finding hot spots (Murray & Estivill-Castro, 1998). For example, Schumacher and Leitner (1999) used the k-means algorithm to identify hot spots in the downtown areas of Baltimore. Comparing these for different years, they found evidence of the displacement of crimes following

redevelopment of the downtown area. Corresponding proactive strategies were then suggested on the basis of the patterns found. Although efficient and scalable in comparison to hierarchical clustering algorithms, partitional clustering algorithms usually require the user to predefine the number of clusters to be found. This, however, is not always feasible (Grubesic & Murray, 2001). Accordingly, researchers have tried to use statistical approaches to conduct hot spot analysis or to test the significance of hot spots (Craglia, Haining, & Wiles, 2000). The test statistics G_i (Getis & Ord, 1992; Ord & Getis, 1995) and Moran's I (Moran, 1950), which are used to test the significance of spatial autocorrelation, can be used to detect hot spots. If a variable is correlated with itself through space, it is said to be spatially autocorrelated. For example, Ratchliffe and McCullagh (1999) employed G_i and G_i^* statistics to identify the hot spots of residential burglary and motor vehicle crimes in a city. Compared with a domain expert's perception of the hot spots, this approach was shown to be effective (Ratchliffe & McCullagh, 1999). Statistical approaches have also been used in crime prediction. Based on spatial choice theory (McFadden, 1973), Xue and Brown (2003) modeled the probability of a criminal choosing a target location as a function of multiple spatial characteristics of the location such as family density per unit area and distance to highway. Using regression analysis, they predicted the locations of future crimes in a city. Evaluation showed that their models significantly outperformed conventional hot spot models. Similarly, Brown, Dalton, and Hoyle (2004) built a logistic regression model to predict suicide bombing in counter-terrorism applications.

Commercially available geographical information systems (GIS) and crime mapping tools, such as ArcView and MapInfo, have been widely used in law enforcement and intelligence agencies for analyzing and visualizing spatial patterns of crimes. Geographical coordinate information as well as various spatial features, such as the distance between the location of a crime to major roads and police stations, is often used in GIS (Harris, 1990; Weisburd & McEwen, 1997).

Research on temporal patterns of crimes is relatively scarce in comparison to crime mapping. Two major approaches have been reported, namely *visualization* and *statistical modeling* approaches. Visualization approaches present individual or aggregated temporal features of crimes using a periodic or timeline view. Common methods of viewing periodic data include sequence charts, point charts, bar charts, line charts, and spiral graphs displayed in 2-D or 3-D (Tufte, 1983). In a timeline view, a sequence of events is presented based on its temporal order. For example, LifeLines provides the visualization of a patient's medical history using a timeline view. The Spatial Temporal Visualizer (STV) (Buetow, Chaboya, O'Toole, Cushna, Daspit, Peterson, et al., 2003) seamlessly incorporates periodic view, timeline view, and GIS view in the system to support criminal investigations. Visualization approaches rely on human users to interpret data presentations and to find temporal patterns of events. Statistical approaches, on the other hand, build statistical models from

observations to capture the temporal patterns of events. For instance, Brown and Oxford (2001) developed several statistical models including a log-normal regression model, a Poisson regression model, and cumulative logistic regression models to predict the number of breaking and entering crimes in Richmond, Virginia. The log-normal regression model was found to fit the data best.

Criminal Network Analysis

Criminals seldom operate in a vacuum but instead interact with one another to carry out various illegal activities. Relationships between individual offenders form the basis for organized crime and are essential for the smooth operation of a criminal enterprise (Cronin, 2005; Strickland, 2002a, 2002b, 2002c, 2002d, 2002e). Unlike bureaucratic organizations, criminal enterprises often operate in networks consisting of nodes (individual offenders) and links (relationships). In criminal networks, there may exist groups or teams, within which members have close relationships. One group may also interact with other groups to obtain or transfer illicit goods, services, or information. Moreover, individuals play different roles in their groups. For example, some key members may act as leaders to control the activities of a group, while others may serve as gatekeepers to ensure the smooth flow of information or illicit goods (Strickland, 2002a, 2002b, 2002c, 2002d, 2002e).

Structural network patterns in terms of subgroups, intergroup interactions, and individual roles thus are important for understanding the organization, structure, and operation of criminal enterprises. Such knowledge can help law enforcement and intelligence agencies disrupt criminal networks and develop effective control strategies to combat organized crime (Cronin, 2005). For example, removal of central members in a network may effectively upset the operational network and put a criminal enterprise out of action (Baker & Faulkner, 1993; McAndrew, 1999; Sparrow, 1991). Subgroups and interaction patterns between groups are helpful for finding a network's overall structure, which often reveals points of vulnerability (Evan, 1972; Ronfeldt & Arquilla, 2001). For a centralized structure such as a star or a wheel, the point of vulnerability lies in its central members. A decentralized network such as a chain or clique, however, does not have a single point of vulnerability and thus may be more difficult to disrupt (Strickland, 2002a, 2002b, 2002c, 2002d, 2002e).

Social Network Analysis (SNA) provides a set of measures and approaches for structural network analysis (Wasserman & Faust, 1994). These techniques were originally designed to discover social structures in social networks (Wasserman & Faust, 1994) and are especially appropriate for studying criminal networks (McAndrew, 1999; Sparrow, 1991). Studies involving evidence mapping in fraud and conspiracy cases have employed SNA measures to identify central members in criminal networks (Baker & Faulkner, 1993; Saether & Canter, 2001). In general,

SNA is capable of detecting subgroups, identifying central individuals, discovering between-group interaction patterns, and uncovering a network's overall structure:

- *Subgroup detection.* With networks represented in a matrix format, the matrix permutation approach and cluster analysis have been employed to detect underlying groups that are not otherwise apparent in data (Wasserman & Faust, 1994). Burt (1976) proposed to apply hierarchical clustering methods based on a structural equivalence measure (Lorrain & White, 1971) to partition a social network into positions in which members have similar structural roles. Xu and Chen (2003) employed hierarchical clustering to detect criminal groups in a narcotics network based on the relational strength between criminals.
- *Central member identification.* Centrality deals with the roles of network members. Several measures, such as degree, betweenness, and closeness, are related to centrality (Freeman, 1979). The degree of a particular node is its number of direct links; its betweenness is the number of geodesics (i.e., the shortest paths between any two nodes) passing through it; and its closeness is the sum of all the geodesics between the particular node and every other node in the network. Although these three measures are all intended to illustrate the importance or centrality of a node, they support interpretation of the roles of network members differently. An individual having a high degree measurement, for instance, may be inferred to have a leadership function, whereas an individual with a high level of betweenness may be seen as a gatekeeper in the network. Baker and Faulkner employed these three measures, especially degree, to find the key individuals in a price-fixing conspiracy network in the electrical equipment industry (Baker & Faulkner, 1993). Krebs found that, in the network consisting of the September 11 hijackers (19 in all), Mohamed Atta scored the highest on degree (Krebs, 2001).
- *Discovery of patterns of interaction.* Patterns of interaction between subgroups can be discovered using an SNA approach called blockmodel analysis (Arabie, Boorman, & Levitt, 1978). Given a partitioned network, blockmodel analysis determines the presence or absence of an association between a pair of subgroups by comparing the density of the links between them at

a predefined threshold value. In this way, blockmodeling introduces summarized individual interaction details into interactions between groups so that the overall structure of the network becomes more apparent.

SNA also includes visualization methods that present networks graphically. The Smallest Space Analysis (SSA) approach (Wasserman & Faust, 1994), a branch of Multi-Dimensional Scaling (MDS), is used extensively in SNA to produce two-dimensional representations of social networks. In a graphical portrayal of a network produced by SSA, the stronger the association between two nodes or two groups, the closer they appear on the graph; the weaker the association, the farther apart (McAndrew, 1999). Several network analysis tools, such as Analyst's Notebook (Klerks, 2001), Netmap (Goldberg & Senator, 1998), and Watson (Anderson, Arbetter, Benawides, & Longmore-Etheridge, 1994), can automatically draw a graphical representation of a criminal network. However, these tools do not provide much structural analysis functionality and rely on investigators' manual examinations to extract structural patterns.

The six classes of KDD techniques reviewed here constitute the key components of our proposed ISI research framework. Our focus on the KDD methodology, however, does not exclude other approaches. For example, studies using simulation and multi-agent models have shown promise in the "what-if" analysis of the robustness of terrorist and criminal networks (Carley, Dombroski, Tsvetovat, Reminga, & Kamneva, 2003; Carley, Lee, & Krackhardt, 2002).

In the next section, we present several case studies showing the value and potential of different KDD technologies to accomplish the critical objectives of national security.

ISI in Critical Mission Areas: Case Studies

In response to the challenges of national security, the COPLINK Center at the University of Arizona has developed several research projects to address five of the six critical mission areas identified in the National Strategy for Homeland Security report (U.S. Office of Homeland Security, 2002): intelligence and warning, border and transportation security, domestic counter-terrorism, protecting critical infrastructure and key assets, and emergency preparedness and response. The center's main goal is to develop information and knowledge management technologies appropriate for capturing, accessing, analyzing, visualizing, and sharing law enforcement and intelligence-related information (Chen, Zeng, Atabakhsh, Wyzga, & Schroeder, 2003). Through the following eight case studies, we demonstrate how critical mission issues could be addressed using the knowledge discovery approach. For each case study, we discuss its relevance to national security missions,

data characteristics, technology used, and selected evaluation results. Quantitative studies focused primarily on the performance of the techniques in terms of effectiveness, accuracy, efficiency, usefulness, and so forth. In qualitative studies where quantitative results are not yet available, we summarize and report comments and feedback from our domain experts.

Intelligence and Warning

Although terrorism depends on surprise (U.S. Office of Homeland Security, 2002), terrorist attacks are not random but require careful planning, preparation, and cooperation before execution. To avoid being preempted by authorities, terrorists may disguise their true identities or hide their illegal objectives and intents behind legal activities. Similarly, criminals may try to minimize the possibility of being identified and captured by using falsified identities. To detect hidden intent and potential for future attacks or offenses is the main goal of intelligence and warning systems. In this section, we present two case studies addressing intelligence and warning needs. The first helped to detect deceptive identity records in police data (Wang, Chen, et al., 2004), while in the second, we present our design for an intelligence Web portal to help trace and monitor the Web sites of terrorist organizations (Chen, Qin, Reid, Chung, Zhou, Xi, et al., in press; Reid, Qin, Chung, Xu, Zhou, Schumaker, 2004).

Case Study 1: Detecting Deceptive Criminal Identities

It is common practice for criminals to lie about the particulars of their identities, such as name, date of birth, address, and social security number, in order to deceive police investigators. Inability to validate identity can be used as a warning mechanism because the deception signals an intent to commit future offenses. In this case study, we focus on uncovering patterns of criminal identity deception based on actual criminal records and suggest an algorithmic approach to revealing false identities (Wang, Chen, et al., 2004).

Data used in this study were authoritative criminal identity records obtained from the Tucson Police Department (TPD). These records were structured database entries containing criminal identity information, such as name, date of birth (DOB), address, identification number (e.g., social security number), race, weight, and height. The total number of criminal identity records stored in the TPD databases was over 1.5 million. In order to study the patterns of criminal identity deception, we selected from the TPD database 372 records involving 24 criminals, each having one real identity record and several deceptive records. These sets of deceptive records were not randomly sampled from the database, but were manually extracted by a police detective expert who has served in law enforcement for 30 years. The expert used convenience sampling, in which he reviewed the list of all identity records and chose the deceptive

identity records that he encountered. Because deceptive identities are sparsely distributed in the criminal database, convenience sampling is more effective than random sampling for experimental purposes. As a result, the conclusions may not be statistically valid.

We carefully examined these 372 records and found that deception occurred most often in specific attributes: name, address, birth date, and Social Security Number (SSN). The identity deception patterns in this dataset are shown in Figure 6.2. Name deception, occurring in most cases, includes giving a false first name and a true last name or vice versa, changing the middle initial, and giving a name pronounced similarly but spelled differently. Deception on DOB can consist of, for example, switching places between the month of birth and the day of birth. Similarly, ID deception is often made by changing a few digits of an SSN or by switching their places. In residency deception, criminals usually change only one portion of the address. For example, we found that, in about 87 percent of cases, criminals provided a false street number along with the true street direction, street name, or street type.

To detect deceptive identity records automatically, we employed a similarity-based association mining method to extract associated (similar) record pairs. Based on the deception patterns found, we selected four attributes (name, DOB, SSN, and address) for our analysis. We compared and calculated the similarity between the values of corresponding attributes of each pair of records. If two records were significantly similar, we assumed that at least one of them was deceptive.

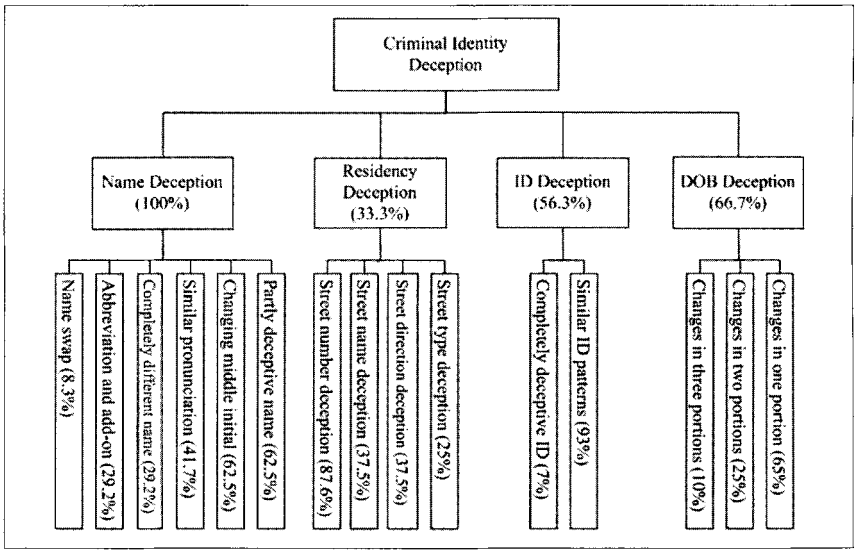

Figure 6.2 Identity deception patterns (each percentage number represents the proportion of records that contain the particular type of deception in the selected dataset).

Because the four selected attributes primarily have string values, we compared two attribute values based on their edit distance (Levenshtein, 1966) and Soundex code (Newcombe et al., 1959). The edit distance between two strings is the minimum number of single character insertions, deletions, and substitutions required to transform one string into the other. Soundex code represents the phonetic pattern of a string. For example, "PEARSE" and "PIERCE" are both coded as "P620." To detect both spelling and phonetic variations between two name strings, edit distance similarity and Soundex similarity were computed separately. In order to capture name exchange deception, similarities were also computed based on different sequences of first name and last name. We took the similarity value from the sequence that had the maximal value between two names. We used only edit distance to compare non-phonetic attributes of DOB, SSN, and address. Each similarity value was normalized between 0 and 1. The similarity value over all four attributes was calculated by means of a normalized Euclidean distance function.

In order to test the performance of our approach, we used convenience sampling again to select another set of 120 records. However in this case, we chose only records with complete information in the name, address, DOB, and SSN fields. The 120 records involved 44 criminals, each of whom had an average of three records in the sample set. Some data were used to train and test our algorithm so that records pointing to the same suspect could be associated with each other. Training and testing were validated by a standard hold-out sampling method. Of the 120 records in the testbed, 80 (66.7 percent) were used for training the algorithm, and the remaining 40 were used for testing.

A similarity matrix was built for all training records. Similarity values in the matrix were used to establish the threshold values appropriate to distinguish between similar and dissimilar pairs. Accuracy rates for correctly recognized similar pairs of records using different threshold values are shown in Table 6.3. When the threshold similarity value was set to 0.52, our algorithm achieved its highest accuracy of 97.4 percent,

Table 6.3 Accuracy comparison based on different threshold values

Threshold	Accuracy	False Negative*	False Positive**
0.6	76.60%	23.40%	0.00%
0.55	92.20%	7.80%	0.00%
0.54	93.50%	6.50%	2.60%
0.53	96.10%	3.90%	2.60%
0.52	**97.40%**	**2.60%**	**2.60%**
0.51	97.40%	2.60%	6.50%
0.5	97.40%	2.60%	11.70%

*False negative: consider disimilar records as similar ones
**False positive: consider similar records as disimilar ones

with relatively small false negative and false positive rates; both were 2.6 percent.

A similarity matrix was also built for the 40 test records. By application of the optimal threshold value to the testing similarity matrix, records having a similarity value of more than 0.52 were considered to be pointing to the same offender. The accuracy of association in the testing data set is shown in Table 6.4. The result shows that the algorithm is effective (with an accuracy level of 94 percent) in linking deceptive records pointing to the same offender.

Although the case study produced promising results, much more research is needed for deception detection, which we believe is a unique and critical problem for ISI.

Table 6.4 The accuracy of association in the testing data set

Threshold	Accuracy	False Negative	False Positive
0.52	94.0%	6.0%	0.0%

Case Study 2: The "Dark Web" Portal

Because the Internet has become a global platform for information dissemination and communication, terrorists also take advantage of the freedom of cyberspace and construct their own Web sites to propagate terrorist ideology, share information, and recruit new members. Web sites of terrorist organizations may also connect to one another through hyperlinks, forming a "dark Web." We are building an intelligent Web portal, called the Dark Web Portal, to help terrorism researchers collect, access, analyze, and understand terrorist groups (Chen, Qin, et al., in press; Reid et al., 2004). This project consists of three major components: Dark Web testbed building, Dark Web link analysis, and Dark Web Portal building.

- *Dark Web Testbed Building.* Drawing on reliable governmental sources such as the Anti-Defamation League (ADL), FBI, and United States Committee for a Free Lebanon (USCFL), we identified 224 U.S. domestic terrorist groups and 440 international terrorist groups. For U.S. domestic groups, group-generated URLs can be found in FBI reports and the Google Directory. For international groups, we used the group names as queries to search major search engines such as Google and manually identified the group-created URLs from the result lists. To ensure that our testbed covered all the major regions in the

world, we sought the assistance of language experts in English, Arabic, Spanish, and Japanese to help us collect URLs in different regions. All URLs collected were manually checked by experts to make sure that they were created by terrorist groups. Once a group's URL was identified, we used the SpidersRUs toolkit, a multilingual Digital Library building tool developed by our own group, to collect all the Web pages under that URL and store them in our testbed. We have collected 500,000 Web pages created by U.S. domestic groups, 400,000 Web pages created by Arabic-speaking groups, 100,000 Web pages created by Spanish-speaking groups, and 2,200 Web pages created by Japanese-speaking groups. This testbed is updated bimonthly.

- *Dark Web Link Analysis and Visualization.* Terrorist groups are not atomized individuals but actors linked to each other through complex networks of direct or mediated exchanges. Identifying how relationships between groups are formed and dissolved in the terrorist group network would enable us to reveal the social milieux and communication channels among terrorist groups across different jurisdictions. Previous studies have shown that the link structure of the Web represents a considerable amount of latent human annotation (Gibson, Kleinberg, & Raghavan, 1998). Thus, by analyzing and visualizing hyperlink structures between terrorist-generated Web sites and their content, we could discover the structure and organization of terrorist group networks, capture network dynamics, and understand their emerging activities (e.g., exploiting formal or informal banking systems, changing identities to take on characteristics more identifiable with Western societies, or creating their own online communities). To test our ideas, we conducted an experiment in which we analyzed and visualized the hyperlink structure between approximately 100,000 Web pages from 46 Web sites in our current testbed. These 46 Web sites were created by four major Arabic-speaking terrorist groups, namely Al-Gama's al-Islamiyya (Islamic Group, IG), Hizballa (Party of God), Al-Jihad (Egyptian Islamic Jihad), and Palestinian Islamic Jihad (PIJ) and their supporters. Hyperlinks between each pair of the 46 Web sites were extracted from the Web pages and a closeness value was calculated for each pair of the 46 Web sites as shown in Figure 6.3. Each node represents a Web site

created by one of the 46 groups. A link existing between two nodes means there are hyperlinks between the Web pages of the two sites. We presented this network to several domain experts and confirmed that the structure of the diagram matched the experts' knowledge of how the groups related to each other in the real world. The four clusters represent a logical mapping of the existing relations among the 46 groups. For instance, the Palestinian terrorist group's cluster includes many of these groups' Web sites, as well as their leaders' sites. Examples include the Al-Aqsa Martyrs' Brigade (http://www.katae.baqsa.org), HAMAS (http://www.ezzedeen.net), and PIJ (http://www.abrarway.com).

- *Dark Web Portal Building*. Using the Dark Web Portal, experts are able to locate specific dark Web information in the testbed quickly through keyword search. To address the information overload problem, the Dark Web Portal is designed with post-retrieval components. A modified version of a text summarizer called TXTRACTOR, which uses sentence-selection heuristics to rank and select important text segments (McDonald & Chen, 2002), has been incorporated into the Dark Web Portal. The summarizer can flexibly summarize Web pages so that experts can quickly get the main idea of a page without having to read through it. A categorizer organizes the search results into various folders labeled with the key phrases extracted by the Arizona Noun Phraser (AZNP) (Tolle & Chen, 2000) from the page summaries or titles, thereby facilitating the understanding of different groups of Web pages. A visualizer clusters Web pages into colored regions using the SOM algorithm (Kohonen, 1995), thus reducing information overload when a large number of search results is obtained. Post-retrieval analysis could further reduce information overload, but researchers are limited to data in their native languages and cannot fully utilize the multilingual information in the testbed. To address this problem, we have added a cross-lingual information retrieval (CLIR) component into the portal. On the basis of our previous research, we have developed a dictionary-based CLIR system for use in the Dark Web Portal. It currently accepts English queries and retrieves documents in English, Spanish, Chinese, Japanese, and Arabic. A machine translation

(MT) component will be added to the Dark Web Portal to translate the multilingual information retrieved by the CLIR component back into the experts' native languages.

Because terrorist groups continue to use the Internet as a communication, recruiting, and propaganda tool, a systematic and system-aided approach to studying their presence on the Web is critically needed.

Border and Transportation Security

Terrorists enter a targeted country by air, land, or sea. The government can improve its counter-terrorism and crime-fighting capabilities by creating a "smart border," where information from borders, customs, transportation, and local law enforcement agencies is integrated and analyzed to help locate wanted terrorists or criminals. Our "BorderSafe" project for cross-jurisdictional information integration and sharing (Marshall, Kaza, Xu, Atabakhsh, Petersen, Violette, et al., in press) illustrates how a smart and safe border might be created.

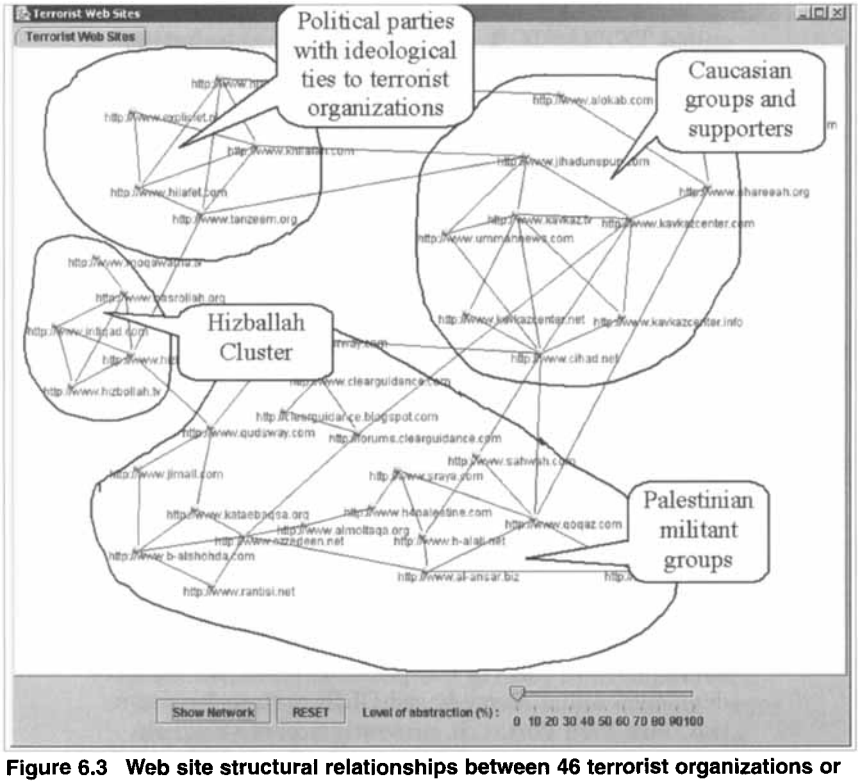

Figure 6.3 Web site structural relationships between 46 terrorist organizations or affiliated groups.

Case Study 3: Enhancing BorderSafe

The BorderSafe project is a collaborative research effort involving the University of Arizona's Artificial Intelligence Lab; several law enforcement agencies including the Tucson Police Department (TPD), Phoenix Police Department (PPD), Pima County Sheriff's Office (PCSO), and Tucson Customs and Border Protection (CBP); the San Diego ARJIS (Automated Regional Justice Systems, a regional consortium of more than 50 public safety agencies); the San Diego Supercomputer Center (SDSC); and the Corporation for National Research Initiatives (CNRI).

In this study our objective was to integrate structured, authoritative data from TPD, PCSO, and a limited dataset from CBP containing license plate data of border crossing vehicles. Tables 6.5 and 6.6 present the statistics from the three datasets. TPD's and PCSO's jurisdictions represent a shared community of citizens in Tucson and southern Arizona. They also share intertwined communities of criminals. We found a substantial amount of data overlap among these datasets. Around seven percent of vehicles involved in gang-related, violent, and narcotics crimes were registered outside of Arizona. More than 483,000 people appeared in both the TPD and PCSO datasets, representing 36 percent of the TPD records and 37 percent of the PCSO records. These statistics strongly suggest that sharing information across jurisdictions could help catch criminals.

The federation approach to data integration was employed. We adopted the COPLINK schema as the global schema and developed a transformation mechanism to reconcile the database structure and semantics from a particular database into the global schema. Data were then mapped or transformed to allow shared query processing. In our

Table 6.5 Statistics regarding the TPD and PCSO datasets

	TPD	PCSO
Number of recorded incidents	2.84 million	2.18 million
Number of persons	1.35 million	1.31 million
Number of vehicles	62,656	520,539

Table 6.6 CBP border crossing dataset

Number of records	1,125,155
Number of distinct vehicles	226,207
Number of plates issued in AZ	130,195
Number of plates issued in CA	5,546
Number of plates issued in Mexico	90,466

datasets, establishing automated transformation procedures for legacy PCSO and TPD records into COPLINK format resolved most of the structural and semantic difference issues.

At the instance level, each dataset had a unique key assigned to each person or vehicle, but these unique keys did not match across datasets. To address this problem, vehicles were matched between datasets on the basis of their license plate numbers. We based people matching on input from domain experts and assumed that all records with the same first name, last name, and DOB represented the same person. These heuristics were not perfect; a few incorrect matches resulted and certainly many correct matches may have been missed. We plan to employ our new identity deception detection approach (Wang, Chen, et al., 2004) in the future to improve instance-level matching.

We generated and visualized several criminal networks based on integrated data. We extracted associations between a set of criminals and vehicles from crime incident records. A link was created when two or more criminals or vehicles were listed in the same incident record. In network visualization we differentiated entity types by shape, key attributes by node color, level of activity (measured by number of crimes committed) by node size, data source by link color, and some details in link text or roll-over tool tips. Figure 6.4 shows a network connecting a known narcotics dealer to a border crossing plate.

A qualitative field study provided positive feedback regarding the potential of our data integration approach. Currently, the crime analysts from both TPD and PCSO are using the triangulated, integrated criminal networks generated by our system to monitor vehicles and criminals crossing the border.

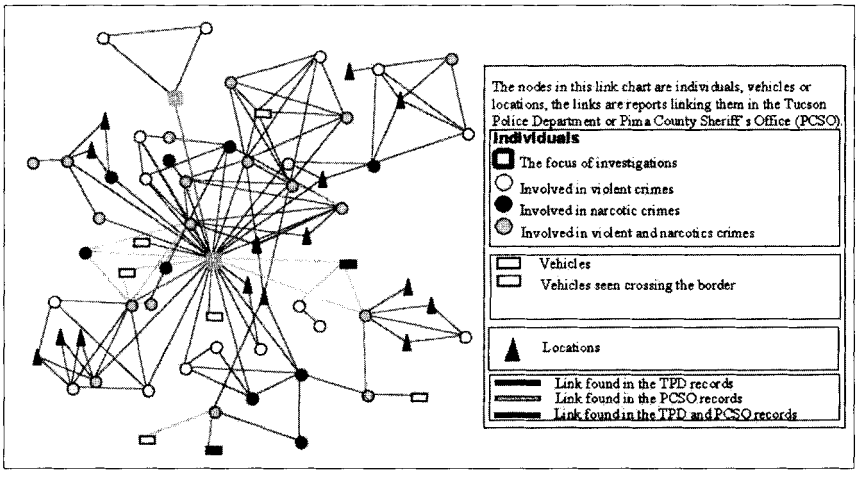

Figure 6.4 A sample criminal network based on integrated data from multiple sources. Nodes and links are color coded in the actual system.

Domestic Counter-Terrorism

As terrorists may be involved in local crimes, state and local law enforcement agencies contribute to national security by investigating and prosecuting these crimes. Terrorism, like gangs and narcotics trafficking, is treated as a type of organized crime in which multiple offenders cooperate to carry out criminal activities. Information technologies that aid in the discovery of cooperative relationships among criminals and reveal the patterns of their interaction would also be helpful in analyzing terrorism. Through three case studies in this section, we show how criminal association information can be extracted from large volumes of data (Hauck et al., 2002) and how structural patterns in criminal or terrorist organizations can be discovered (Xu & Chen, 2003, in press).

Case Study 4: COPLINK Detect

Crime analysts and detectives search for criminal associations to develop investigative leads. However, because association information is not directly available in most existing law enforcement and because intelligence databases and manual searching are extremely time consuming, automatic identification of relationships among criminal entities may significantly speed up investigations. COPLINK Detect is a link analysis system that automatically extracts relationship information from large volumes of crime incident data (Hauck et al., 2002).

Our data were structured crime incident records stored in TPD databases. The TPD's current record management system (RMS) consists of more than 1.5 million crime incident records that contain details of criminal events spanning from 1986 to 2004. Although investigators can access the RMS to tie information together, they must manually search the RMS for connections or existing relationships.

We used the concept space approach (Chen & Lynch, 1992) to identify relationships between entities of interest. Concept space analysis is a type of co-occurrence analysis used in information retrieval. The resulting network-like concept space holds all possible associations between terms—that is, the system retains and ranks every existing link between every pair of concepts. In COPLINK Detect, detailed incident records serve as the underlying space, and concepts are derived from the meaningful terms that occur in each incident. Concept space analysis easily identifies relevant terms and their degree of relationship to the search term. The system output includes relevant terms ranked in the order of their degree of association, thereby distinguishing the most relevant terms from inconsequential ones. From a crime investigation standpoint, concept space analysis can help investigators link known entities to other related entities that might contain useful information for further investigation, such as people and vehicles related to a given suspect. It is considered an example of entity association mining (Lin & Brown, 2003).

Information related to a suspect can move an investigation in the right direction, but revealing relationships among data in one particular incident might fail to capture other relationships from the entire database. In effect, investigators need to review all incident reports related to a suspect and this can be tedious work. The COPLINK Detect system introduces concept space as an alternative method that captures the relationships between four types of entities (person, organization, location, and vehicle) across the entire database. COPLINK Detect also offers an easy-to-use interface and allows searching for relationships among the four types of entities. Figure 6.5 presents the COPLINK Detect interface, showing sample search results for vehicles, relations, and crime case details (Hauck et al., 2002).

We conducted user studies to evaluate the performance and usefulness of COPLINK Detect. Eleven crime analysts and one homicide detective from TPD participated in the longitudinal field study over a four-week period. Crime analysts were experienced in investigating high-profile cases as well as creating statistical reports on criminal activities. They were accustomed to link analysis and are the target user group of COPLINK Detect. Although detectives were not specialized in crime analysis in general, the participating homicide detective was experienced in searching for criminal associations using record management systems. In this study, three major areas were identified where COPLINK Detect provided improved support for crime investigation: link analysis, interface design, and operating efficiency.

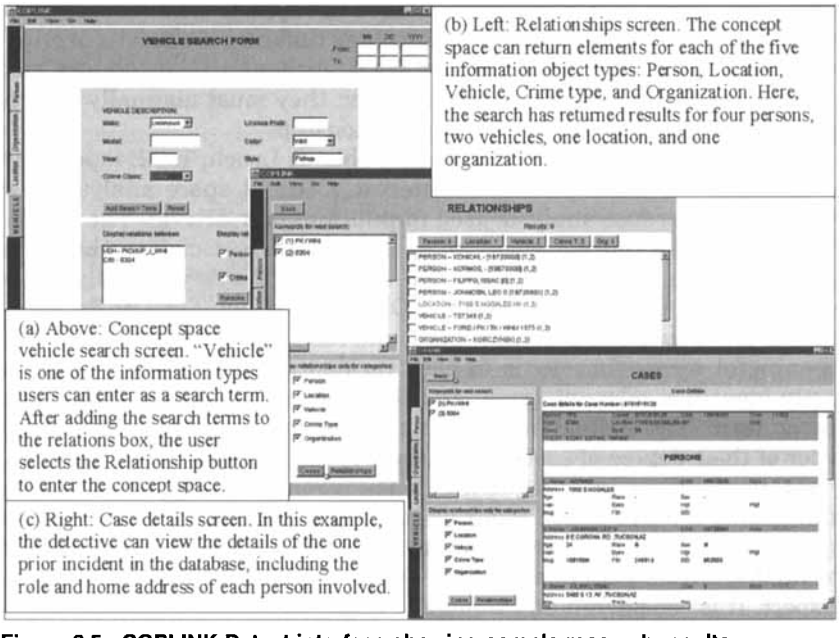

Figure 6.5 COPLINK Detect interface showing sample research results.

Case Study 5: Criminal Network Mining

Because organized crime is carried out by networked offenders, investigation naturally depends on network analysis approaches. Grounded in social network analysis methodology, our criminal network-structure mining research aims at helping intelligence and security agencies extract valuable knowledge regarding criminal or terrorist organizations by identifying the central members, subgroups, and overall network structure (Xu & Chen, 2003, in press).

Two datasets from TPD were used in the study. (1) A gang network: The list of gang members consisted of 16 offenders who had been under investigation during the first quarter of 2002. These gang members had been involved in 72 crime incidents of various types (e.g., theft, burglary, aggravated assault, drug offenses) since 1985. We used the concept space approach and generated links between criminals who had committed crimes together, ending with a network of 164 members. (2) A narcotics network: The list for the narcotics network consisted of 71 criminal names. A sergeant from the Gang Unit had been studying the activities of these criminals since 1995. Because most of them had committed crimes related to methamphetamines, the sergeant called this network the "Meth World." These offenders had been involved in 1,206 incidents since 1983. A network of 744 members was generated.

We made use of SNA approaches to extract structural patterns in the criminal networks:

- *Network partition.* We employed hierarchical clustering, namely the complete-link algorithm, to partition a network into subgroups based on relational strength. Clusters obtained represent subgroups. To employ the algorithm, we first transformed co-occurrence weights generated in the previous phrase into distances/dissimilarities. The distance between two clusters was defined as the distance between the pair of nodes drawn from each cluster that was farthest apart. The algorithm worked by merging the two nearest clusters into one cluster at each step and eventually formed a cluster hierarchy. The resulting cluster hierarchy specified groupings of network members at different granularity levels. At lower levels of the hierarchy, clusters (subgroups) tended to be smaller and group members were more closely related. At higher levels of the hierarchy, subgroups are large and group members may be loosely related.
- *Centrality measures.* We used all three centrality measures to identify central members in a given subgroup. The degree of a node could be obtained by

counting the total number of links it had to all the other group members. A node's score of betweenness and closeness required the computation of shortest paths (geodesics) using Dijkstra's (1959) algorithm.

- *Blockmodeling.* At a given level of a cluster hierarchy, we compared intergroup link densities with the network's overall link density to determine the presence or absence of intergroup relationships.
- *Visualization.* To map a criminal network onto a two-dimensional display, we employed multi-dimensional scaling (MDS) to generate x-y coordinates for each member in a network. We chose Torgerson's (1952) classical metric MDS algorithm because

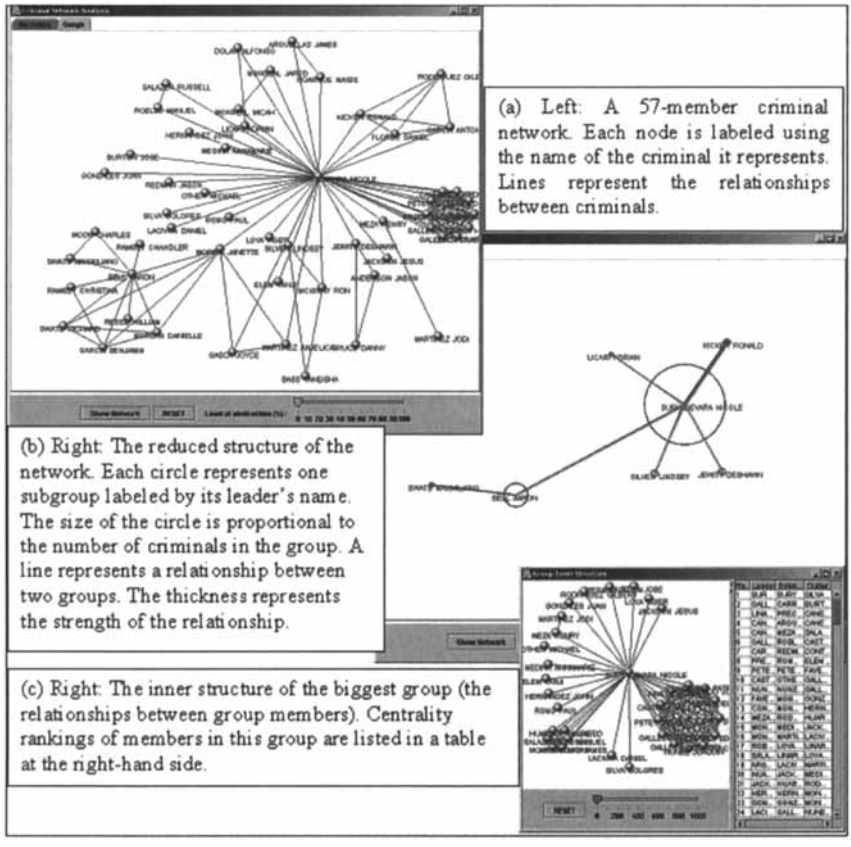

Figure 6.6 An SNA-based system for criminal network analysis and visualization.

distances transformed from co-occurrence weights were quantitative data.

A graphical user interface was provided to visualize criminal networks. Figure 6.6 shows the screenshot of our prototype system. In this example, each node was labeled with the name of the criminal it represented. Criminal names were scrubbed for data confidentiality. A straight line connecting two nodes indicated that two corresponding criminals committed crimes together and thus were related. To find subgroups and interaction patterns between groups, a user could adjust the "level of abstraction" slider at the bottom of the panel. A high level of abstraction corresponded with a high distance level in the cluster hierarchy. Group members' rankings in centrality are listed in a table.

A qualitative study was conducted to evaluate the prototype system. We presented the two testing networks to domain experts at TPD and received encouraging feedback (Xu & Chen, 2003):

- *Subgroups detected were mostly correct.* The domain experts checked and validated the members in each group. These groups had different characteristics with different specialties or crime preferences. We also found that although relationships in our network were extracted based on crime incidents, they reflected relationships between criminals based on friendship, kinship, and even conflicts.
- *Centrality measures provided ways of identifying key members in a network.* According to our domain experts, betweenness was a reliable measure to identify gatekeepers between subgroups. However, degree sometimes misidentified leaders because the criminals with the most connections to others may not always be the leaders. Leaders may be smart enough to hide behind other criminals to avoid police contact.
- *Interaction patterns identified could help reveal relationships that previously had been overlooked.* Our system could generate the "big picture" for a complex network. As a result some relationships between criminal groups that had been overlooked before the system were made easier to identify.
- *Saving investigation time.* Our domain experts had obtained knowledge about the gang and narcotics organizations based on several years of work. Using information gathered from a large number of arrests and interviews, they had built the networks incrementally by linking new criminals to known

gangs in the network and then studying the organization of these networks. Because there was no structural analysis tool available, they did all of this by hand. With the help of our system, they expected that substantial time would be saved in network creation and structural analysis.

- *Saving training time for new investigators.* New investigators who did not have sufficient knowledge of criminal organizations and individuals could use the system to grasp the essence of the network and related crime history quickly. They would not have to spend as much time studying hundreds of incident reports.
- *Helping prove guilt of criminals in court.* The relationships discovered between individual criminals and criminal groups would be helpful for proving guilt when presented at court for prosecution.

Case Study 6: Analyzing Terrorist Networks

As part of the worldwide Islamic Jihadist movement, a number of terrorist organizations have targeted the West. Terrorism and terrorist attacks pose severe threats and have caused significant damage worldwide. Only with an in-depth understanding of terrorism and terrorist organizations can societies defend themselves against the threats. Because terrorist organizations often operate in networks through which individual terrorists collaborate to carry out attacks (Klerks, 2001; Krebs, 2001), network analysis can help uncover valuable information by studying the networks' structural properties (Xu & Chen, in press). We have employed techniques and methods from SNA and Web mining to address the problem of structural analysis of terrorist networks.

The objective of this case study was to examine the potential of network analysis tools for terrorist analysis. By comparing our findings with experts' input we sought to ascertain whether automatic analysis of structural properties of a terrorist network would generate information consistent with expert knowledge.

In this study, we focused on the structural properties of a set of Islamic terrorist networks, including Osama bin Laden's Al Qaeda. In a recently published book, Sageman (2004) documented the history and evolution of these terrorist organizations, which he terms Global Salafi Jihad (GSJ). Sageman is a social psychologist and formerly served as a foreign service officer. During the Afghan-Soviet war, from 1986 to 1989, he dealt with Islamic fundamentalists on a daily basis and developed substantial expertise in terrorism and terrorist organizations. Drawing upon various open sources, such as news articles and court transcripts, he collected data on

364 terrorists in the GSJ network regarding their background, religious beliefs, social relations, and the terrorist attacks in which they participated. There are three types of social relations among these terrorists: personal links (e.g., acquaintance, friendship, and kinship), operational links (e.g., collaborators in the same attack), and relations formed after attacks (Sageman, 2004). Sageman identified four major terrorist groups on the basis of their geographical locations: Central Staff, Core Arab, Maghreb Arab, and Southeast Asian. Each group has its own leaders. For example, Osama bin Laden is the leader of the Central Staff group, which connects to the other three groups through several lieutenants.

We analyzed the GSJ network based on the social relation data contained in a spreadsheet provided by Sageman. Using the SNA visualization approach, we depicted the GSJ network graphically as shown in Figure 6.7.

- *Centrality analysis.* Considering all three types of social relations, we found that the four group leaders were among the 11 most popular members, where popularity was represented by degree measure. For

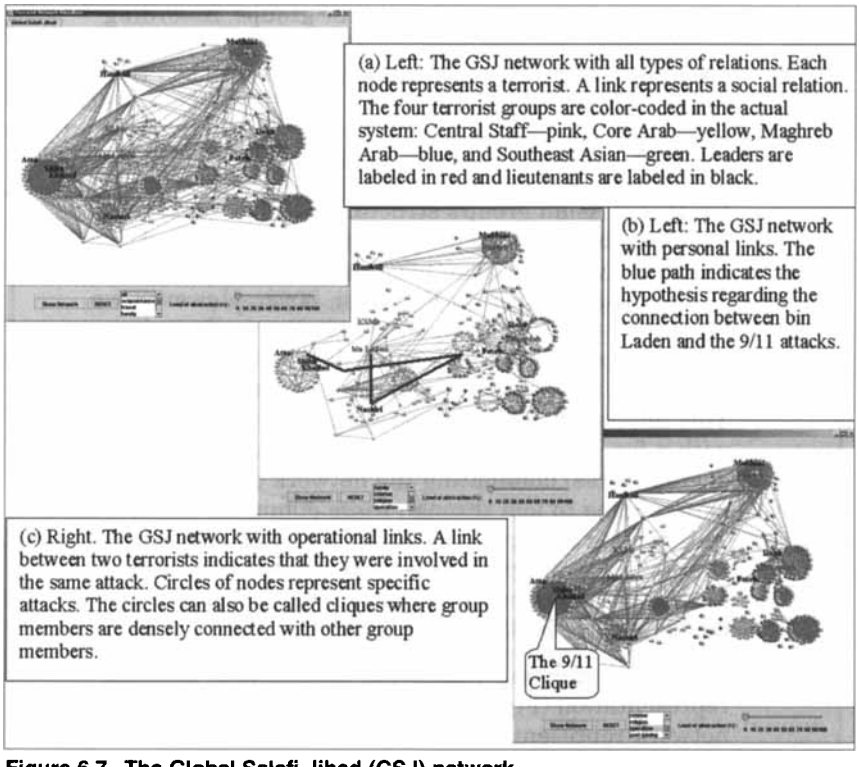

Figure 6.7 The Global Salafi Jihad (GSJ) network.

example, Osama bin Laden had 72 links to other terrorists and ranked second in degree. Although he was not a leader, Hambali had the highest degree score and played an important role in connecting different terrorist groups (see Figure 6.7a). Moreover, the lieutenants tended to have high scores in betweenness and served as gatekeepers between groups. The analysis implies that centrality measures could be useful for identifying important members of a terrorist network.

- *Subgroup analysis.* The four terrorist groups depicted in Figure 6.7 were color coded in the actual GSJ network system using Sageman's advice. To find out whether these geographically based groups were also structurally cohesive, we calculated the cohesion score (Wasserman & Faust, 1994) of each group. We found that all these groups had high cohesion scores. The Southeast Asian group scored the highest in cohesion. This may suggest that members in this group tended to be more closely related to members of their own group than to members from other groups. According to Sageman, the Southeast Asian group was quite different from the other three groups in terms of their religious beliefs and missions.
- *Network structure analysis.* Sageman had reported that these groups had different structures: The Southeast Asian group's structure was hierarchical with members at higher levels leading lower-level members, whereas the other three groups were scale-free networks (Albert & Barabási, 2002). However, we found that the four groups were similar in their degree distribution, which was a power-law distribution with a long tail for large values of degree (see Figure 6.8). This implies that all four networks were scale-free, with a few important members (nodes with high degree scores) dominating the network and new members tending to join through these dominant members. This finding has an important policy implication: Disruptive strategies should potentially be focused on central members in a terrorist network (Strickland, 2002a, 2002b, 2002c, 2002d, 2002e).
- *Link path analysis.* Comparing the personal network representation (Figure 6.7b) and the operational network representation (Figure 6.7c), we found that some important members did not have direct personal

links to an attack prior to execution. For example, neither Osama bin Laden, Khalid Sheikh Mohammed, nor Hambali had direct personal links to terrorists in the 9/11 attack clique. We performed link path analysis to find out the shortest paths of personal links leading to the 9/11 terrorists. One of our hypotheses was that Osama bin Laden connected to the 9/11 clique through a four-hop path: bin Laden—Nashiri—ZaMihd—Mihdhar—Shibh (the dark path in Figure 6.7b). Although this hypothesis turned out to be wrong according to Sageman's feedback (other information was needed to establish the link), the analysis showed the potential of using link path analysis to generate hypotheses about the motives and planning processes behind terrorist attacks.

Protecting Critical Infrastructure and Key Assets

The Internet is a critical infrastructure and asset in the information age. Cybercriminals have been using various Web-based channels (e.g., e-mail, Web sites, Internet newsgroups, chat rooms) to distribute illegal materials. One common characteristic of these channels is anonymity. People usually do not need to provide information about their real identity, such as name, age, gender, and address, in order to participate in cyberactivities. Compared with conventional crimes, cybercrime conducted through such anonymous channels creates novel challenges for

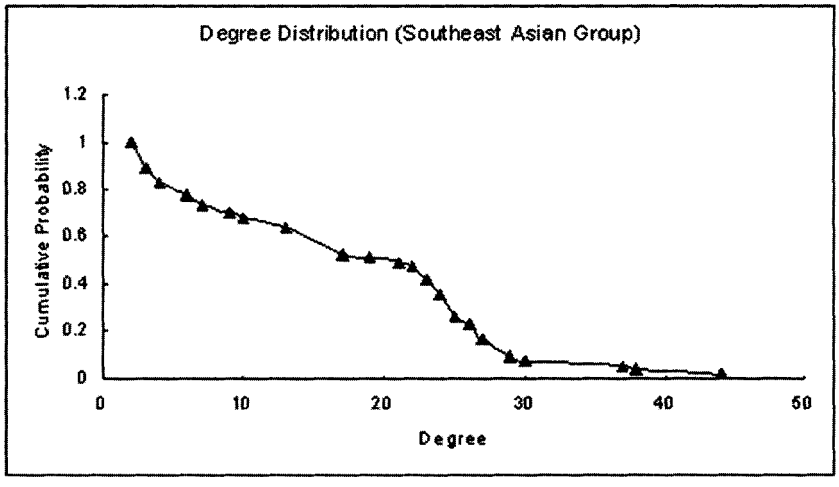

Figure 6.8 The power-law degree distribution of the Southeast Asian group.

researchers and law enforcement agencies engaged in criminal identity tracing. The situation is further complicated by the enormous number of cyberusers and activities, making the manual approach to criminal identity tracing impossible. Law enforcement agencies urgently need approaches that automate criminal identity tracing in cyberspace and allow investigators to prioritize their tasks and focus on major criminals. This case study demonstrates the potential of using authorship analysis with carefully selected feature sets and effective classification techniques for criminal identity tracing in cyberspace (Zheng et al., 2003).

Case Study 7: Identity Tracing in Cyberspace

Data used in this study were from open sources. Three datasets, two in English and one in Chinese, were collected. One of the English datasets consisted of 153 Usenet newsgroup illegal sales of pirated CDs and software messages. We manually identified the nine most active users (represented by a unique ID and e-mail address) who posted messages in these newsgroups. The Chinese dataset contained 70 Bulletin Board System (BBS) illegal CD and software for-sale messages downloaded from a popular Chinese BBS.

The two key techniques used in this study were feature selection and classification. The objective was to classify text messages into different classes with each class representing one author. Based on a review of previous studies on text and e-mail authorship analysis, along with the specific characteristics of the messages in our datasets, we selected a large number of features that were potentially useful for identifying message authors. Three types of features were used: *style markers* (content-free features such as frequency of function word, total number of punctuation marks, and average sentence length), *structural features* (such as use of a greeting statement, position of requoted text, use of farewell statement), and *content-specific features* (such as frequency of keywords, special character of content).

For classification analysis, three popular classifiers were selected including the C4.5 decision tree algorithm (Quinlan, 1986), backpropagation neural networks (Lippmann, 1987), and support vector machines (SVM) (Cristianini & Shawe-Taylor, 2000; Hsu & Lin, 2002). Each individual classifier had been employed in previous authorship analysis research (Diederich, Kindermann, Leopold, & Paass, 2003). In general, SVM and neural networks had exhibited better performance than decision trees (Diederich et al., 2003). However, most previous authorship studies had been based on corpora of newspaper articles such as *The Federalist Papers*. Because online messages are quite different from formal articles in style, we needed to test the performances of these three algorithms on our datasets.

The procedure of the experiment was as follows: Three experiments were conducted on the newsgroup dataset with one classifier at a time. First, 205 style markers (67 for the Chinese BBS dataset) were used, nine structural features were added in the second run, and nine content-specific

features were added in the third run. A 30-fold cross-validation testing method was used in all experiments.

We used *accuracy*, *recall*, and *precision* to evaluate the prediction performance of the three classifiers. Accuracy represents the overall prediction performance of a classifier. For each author, we used precision and recall to measure the effectiveness of a classifier. The three measures are defined in equations (1)–(3).

$$(1)\quad \text{Accuracy} = \frac{\text{Number of messages with author correctly identified}}{\text{Total number of messages}}$$

$$(2)\quad \text{Precision} = \frac{\text{Number of messages correctly assigned to the author}}{\text{Total number of messages assigned to the author}}$$

$$(3)\quad \text{Recall} = \frac{\text{Number of messages correctly assigned to the author}}{\text{Total number of messages written by the author}}$$

We summarize the results as follows:

- *SVM and neural networks outperformed the C4.5 decision tree algorithm.* For example, in regards to the application of style markers to the e-mail dataset, the C4.5, neural networks, and SVM achieved accuracies of 74.29 percent, 81.11 percent, and 82.86 percent, respectively. SVM also consistently achieved higher accuracy, precision, and recall than the neural networks. However, the performance differences between SVM and neural networks were relatively small. Our results were generally consistent with previous studies, in that neural networks and SVM typically achieve better performance than decision tree algorithms (Diederich et al., 2003).
- *Use of style markers and structural features outperformed use of style markers only.* We achieved significantly higher accuracy levels for all three datasets (p-values were below 0.05) by adopting the structural features. This possibly resulted from an author's consistent writing patterns being evident in the message's structural features.
- *Use of style markers, structural features, and content-specific features did not achieve better performance than use of style markers and structural features.* The results indicated that using content-specific features as additional features did not improve the authorship prediction performance significantly (with p-value of 0.3086). We thought this was because

authors of illegal messages typically included diverse content in their messages and little additional information could be derived from the message content to determine authorship. We also observed that high levels of accuracy were obtained when style markers alone were used as input features for the English datasets. The accuracy level ranged from 71 to 89 percent. The results indicated that style markers alone contain a large amount of information about people's online message writing styles and are surprisingly robust in predicting the authorship.

- *There was a significant drop in prediction performance measures for the Chinese BBS dataset in comparison to the English datasets.* For example, when using style markers only, C4.5 achieved average accuracies of 86.28 and 74.29 percent for the English newsgroup and e-mail datasets, whereas for the Chinese dataset, it achieved an average accuracy of only 54.83 percent. A possible reason was that only 67 Chinese style markers were used in the experiments, significantly fewer than the 205 style markers used with the English dataset. We expect to achieve higher prediction performances if additional Chinese style markers are identified and included. We also observed that when structural features were added, all three algorithms achieved relatively high precision, recall, and accuracy (from 71 to 83 percent) for the Chinese dataset. Considering the significant language differences, our proposed approach to the problem of online message identity tracing appears promising in a multilingual context.

Similar to "finger-print" and "voice-print" that could help identify a person, we believe that there is a need and potential for developing a robust multilingual "write-print" model based on an individual's unique writing style. Such a model, possibly building on research in stylometrics (Williams, 1975) would have strong value for cybercrime investigation.

Emergency Preparedness and Responses

Terrorist attacks can cause devastating damage to a society through the use of chemical, biological, or radiological weapons. Currently, a large amount of infectious disease data is being collected by various laboratories, health care providers, and government agencies at local, state, national, and international levels (Pinner, Rebmann, Schuchat, & Hughes, 2003). However, access to some of these data sources and related search and reporting functionalities may be limited to the agencies that

have developed such systems (Kay, Timperi, Morse, Forslund, McGowan, & O'Brien, 1998), reducing the effective use of infectious disease data in national and global contexts. In addition, real-time data sharing, especially of databases across species and jurisdictions, could enhance expert scientific review and rapid response using input and action triggers provided by multiple government and public health partners. In this case study we discuss our ongoing research and system development efforts designed to address some of these challenges. We aim to develop scalable technologies and related standards and protocols needed for a national infectious disease information infrastructure (Zeng, Chen, Tseng, Larson, Eidson, Gotham, et al., 2004).

Case Study 8: The WNV-BOT Portal

Our research focuses on two prominent infectious diseases: *West Nile Virus* (WNV) and *Botulism*. These two diseases were chosen because of their significant public health and national security implications and the availability of related datasets for the states of New York and California. We developed a research prototype called the WNV-BOT Portal system, which provides integrated, Web-enabled access to a variety of distributed data sources including the New York State Department of Health (NYSDH), the California Department of Health Services (CADHS), and other federal sources (e.g., the United States Geological Survey [USGS]). It also provides advanced information visualization capabilities as well as predictive modeling support.

Architecturally, the WNV-BOT Portal consists of three major components: a *Web portal*, a *data store*, and a *communication backbone*. The Web portal implements the user interface and provides the following main functionalities: (1) searching and querying available WNV/BOT datasets, (2) visualizing WNV/BOT datasets using spatial-temporal visualization, (3) accessing analysis and prediction functions, and (4) accessing the alerting mechanism.

To enable data interoperability, we use Health Level Seven (HL7) standards (http://www.hl7.org) as the main storage format. In our data warehousing approach, contributing data providers transmit data to WNV-BOT Portal as HL7-compliant XML messages (through a secure network connection if necessary). After receiving these XML messages, the WNV-BOT Portal adds them directly to its data store. To alleviate potential computational performance problems associated with this HL7 XML-based approach, we have identified a core set of data fields, on which searches could be performed efficiently.

An important function of the data store layer is data ingest and access control. The data ingest control module is responsible for checking the integrity and authenticity of data feeds from the underlying information sources. The access control module is responsible for granting and restricting user access to sensitive data.

The communication backbone component enables data exchanges between the WNV-BOT Portal and the underlying WNV/BOT sources

based upon the CDC's (Centers for Disease Control and Prevention) Electronic Disease Surveillance System (NEDSS) and HL7 standards. It uses a collection of source-specific "connectors" to communicate with underlying sources. We use the connector linking NYSDOH's Health Information Network (HIN) system and WNV-BOT Portal to illustrate a typical design of such connectors. The data sent from HIN to the portal system are transmitted in a "push" manner. HIN sends secure Public Health Information Network Messaging System (PHIN MS) messages to the portal at prespecified time intervals. The connector at the portal side runs a data receiver daemon listening for incoming messages. After a message is received, the connector checks for data integrity syntactically and invokes the data normalization subroutine. Then the connector stores the verified message in the portal's internal data store through its data ingest control module. Other data sources (e.g., those from USGS) may have "pull-" type connectors, which periodically download information from the source Web sites and examine and store data in the portal's internal data store. In general, the communication backbone component provides data receiving and sending functionalities, source-specific data normalization, as well as data encryption capabilities.

The WNV-BOT Portal makes available the Spatial Temporal Visualizer (STV) (Buetow et al., 2003) to facilitate exploration of infectious disease case data and to summarize query results. STV has three integrated and synchronized views: periodic, timeline, and GIS. Figure

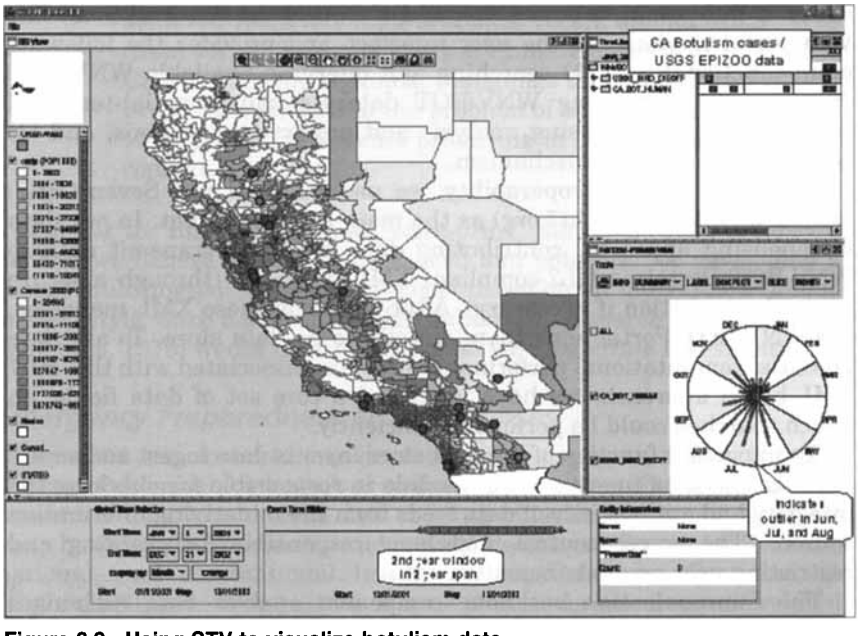

Figure 6.9 Using STV to visualize botulism data.

6.9 illustrates how these three views can be used to explore the infectious disease dataset. The top-left panel shows the GIS view. The user can select multiple datasets to be shown on the map in a layered manner using the checkboxes. The top-right panel corresponds to the timeline view displaying the occurrences of various cases using a Gantt chart-like display. The user can also access case details easily by using the tree display located left of the timeline display. Below the timeline view is the periodic view through which the user can identify periodic temporal patterns (e.g., which months have an unusually high number of cases). The bottom portion of the interface allows the user to specify subsets of data to be displayed and analyzed.

Our project has supported exploration of, and experimentation with, technological infrastructures needed for a full-fledged implementation of a national infectious disease information infrastructure and has helped foster information sharing and collaboration among related government agencies at state and federal levels. In addition, we have obtained important insights into, and hands-on experience with, various important policy-related challenges faced by developing a national infrastructure. For example, a nontrivial part of our project activity has been devoted to developing data-sharing agreements between project partners from different states.

Our ongoing technical research is focusing on two aspects of infectious disease informatics: hotspot analysis and efficient alerting and dissemination. For WNV, localized clusters of dead birds typically identify high-risk disease areas. Automatic detection of dead bird clusters using hotspot analysis can help predict disease outbreaks and allocate prevention/control resources effectively. Initial experimental results indicate that these techniques are promising for disease informatics analysis. We are planning to augment existing predictive models by considering additional environmental factors (e.g., weather information, bird migration patterns), and tailoring data mining techniques for infectious disease datasets that have prominent temporal features.

Case Study Summary

We summarize in Table 6.7 the eight case studies in terms of their data characteristics, technologies employed, and the national security missions they addressed using our proposed ISI research framework.

The ISI Partnership Framework

In order to accomplish the six critical mission areas of national security, the Department of Homeland Security has proposed establishing a network of laboratories consisting of satellite research centers across the nation (U.S. Office of Homeland Security, 2002). The purpose is to create a multidisciplinary environment for developing technologies to counter various threats to homeland security. However, information sharing and

Table 6.7 Summary of ISI case studies

Case Study	Project	Data Characteristics	Technologies Used	Critical Mission Area Addressed
1	Identity deception detection	• Authoritative source • Structured criminal identity records	• Association mining • Similarity-based	Intelligence and warning
2	Dark Web Portal	• Open source • Web hyperlink data	• Cluster analysis • Visualization	Intelligence and warning
3	BorderSafe	• Authoritative source • Structured data	• Information sharing and integration • Database federation	Border and transportation security
4	COPLINK Detect	• Authoritative source • Structured data	• Association mining • Statistical-based	Domestic counter-terrorism
5	Criminal network analysis	• Authoritative source • Structured data	• Social network analysis • Cluster analysis • Visualization	Domestic counter-terrorism
6	Terrorist network analysis	• Open source • Text data, structured data	• Intelligence text mining • Social network analysis	Domestic counter-terrorism
7	Identity tracing in cyberspace	• Open source • Structured data	• Intelligence text mining • Classification	Protecting critical infrastructure and key assets
8	WNV-BOT Portal	• Authoritative source • Structured data	• Information sharing and integration • Spatial and temporal visualization	Emergency preparedness and responses

collaboration across different jurisdictions, agencies, and research institutes is not merely a technical issue. A variety of social, organizational, and political barriers needs to addressed, including:

- *Security and confidentiality.* In the intelligence and law enforcement domain, security is of great concern. Data regarding crimes, criminals, terrorist organizations, and potential terrorist attacks may be highly sensitive and confidential in nature. Improper use of data could lead to fatal consequences.
- *Trust and willingness to share information.* Different agencies may not be motivated to share information and collaborate if there is no immediate gain. They may also fear that information being shared will be misused, resulting in legal liabilities.
- *Data ownership and access control.* The questions that need to be addressed are: Who owns a particular data set? Who is allowed to access, aggregate, or input data? Who owns the derivative data (knowledge)? For both original and derivative data, who is allowed to distribute them to whom?

The COPLINK Center at the Artificial Intelligence Lab of the University of Arizona, as a leading research center for law enforcement and intelligence information and knowledge management, intends to become a part of the national network of laboratories. During its development over the past decade the COPLINK Center has encountered many of these non-technical challenges in its partnerships with various law enforcement and federal agencies. In this section, we summarize some of our experiences and lessons learned.

Ensuring Data Security and Confidentiality

In any data sharing initiative, it is essential to make sure that the data shared between agencies are secure and that the privacy of individuals is respected. In our research we have taken the necessary measures to ensure data privacy, security, and confidentiality. Data shared among law enforcement agencies, such as TPD, PPD, and CBP, contained only law enforcement data and were available only to individuals screened by these agencies using a combination of TPD Background Check, Employee Non-Disclosure Agreement (NDA), and the Terminal Operator Certificate (TOC) test.

All personnel who have access to law enforcement data fill out background forms provided by TPD and have their fingerprints taken at TPD. They also sign a nondisclosure agreement provided by TPD. In addition, they take the TOC test every year. The background information and fingerprints are then checked by TPD investigators to ensure lack of involvement in criminal activity and to verify identity.

In addition to these forms and test, all law enforcement data in the University of Arizona COPLINK Center reside behind a firewall and in a secure room accessible only by activated cards to those who have met the security criteria. As soon as an employee stops working on projects related to law enforcement data, his or her card is deactivated. However, the NDA is perpetual and remains in effect even after a researcher or employee leaves. These requirements are similar to those imposed upon noncommissioned civilian personnel in a police department.

Reaching Agreements Among Partners

Federal, state, and local regulations require that agreements between agencies within their respective jurisdictions receive advanced approval from their governing hierarchy. This precludes informal information sharing agreements between those agencies. We found that requirements varied from agency to agency according to the statutes by which they were governed.

For instance, the ordinances governing information sharing by the city of Tucson differed somewhat from those governing the city of Phoenix. This necessitated numerous attempts and passes at proposed documents by each city's law enforcement and legal staffs before a final draft could be settled upon for approval by the city councils. We found that similar

language existed in the ordinances and statutes governing this exchange, but that the processes varied significantly. It appears that the level of bureaucracy is proportional to the size of the jurisdiction.

TPD has recently developed a generic Inter-Governmental Agreement (IGA) that could be adopted between different law enforcement agencies. This IGA was condensed from memoranda of understanding (MOUs), policies, and agreements that previously existed in various forms between numerous agencies. The IGA was drafted to be generic, including language from those laws but excluding reference to any particular chapter or section. This allowed the required verbiage to exist in the document without being specific to any jurisdiction.

Sharing information between agencies with disparate information systems has also led to the bridging of boundaries between software vendors and agencies (their customers). We took care not to violate licensing terms by ensuring that nondisclosure agreements existed and that contract language assured compliance with the vendors' licensing policies.

We believe MOUs and IGAs can be used as templates of information sharing agreements and contracts, and can serve as components of an ISI partnership framework. We plan to provide free access to these legal agreement templates to help facilitate the process of information sharing and collaboration across agencies and research institutions in the future.

Conclusions and Future Directions

In this chapter we have discussed the technical issues related to intelligence and security informatics research, which supports accomplishment of the critical missions of national security. We have proposed a research framework addressing the technical challenges facing counterterrorism and crime-fighting applications, with a primary focus on knowledge discovery from databases (KDD). We have identified and incorporated into the framework six classes of ISI technologies: information sharing and collaboration, crime association mining, crime classification and clustering, intelligence text mining, spatial and temporal analysis of crime patterns, and criminal network analysis. We have also presented a set of COPLINK case studies, ranging from the detection of criminal identity deception to an intelligent Web portal for monitoring terrorist Web sites, thus demonstrating the potential of ISI technologies for contributing to the critical missions of national security.

As this new ISI domain continues to evolve, several important directions need to be pursued, including technology development; testbed creation; and social, organizational, and policy studies:

- New technologies need to be developed and many existing information technologies should be re-examined and adapted for national security applications. The knowledge discovery perspective

provides a promising direction. However, new technologies should be developed in a legal and ethical framework that does not compromise the privacy or civil liberties of private citizens.

- Large scale, nonsensitive data testbeds that incorporate data from diverse, authoritative, and open sources and in different formats should be created and made available to the ISI research community. Lack of real data has been a long-standing problem in intelligence- and security-related research. Many researchers are forced to use simulated or synthetic data that may not resemble actual crime data characteristics. Furthermore, comparing competing technical approaches has been difficult because of the lack of standard test collections. A comprehensive and non-sensitive open source data collection, analogous to the Message Understanding Conference collection, would be of great value for ISI researchers to experiment, test, and evaluate various technologies and to compare and share findings, insights, and knowledge. Advanced methods may need to be employed to scrub data contained in the non-open source testbed to ensure data confidentiality while preserving its characteristics and underlying structures.

The ultimate goal of ISI research is to enhance national security. However, the question of how this type of research has and will have an impact on society, organizations, and the general public remains unanswered. Researchers from sociology, political science, organizational and management sciences, psychology, and education can contribute substantially to this task.

We hope that active ISI research will help improve knowledge discovery and dissemination; enhance information sharing and collaboration among academics, industry, and local, state, and federal agencies; and thereby promote positive societal outcomes.

Acknowledgments

The projects reported in the case studies have been funded mainly by the following grants:

National Institute of Justice (NIJ), COPLINK: Database Integration and Access for a Law Enforcement Intranet, 1997–2000.

National Science Foundation (NSF), Digital Government Program, COPLINK Center: Information and Knowledge Management for Law Enforcement, 2000–2003.

National Science Foundation (NSF) and Central Intelligence Agency (CIA), Knowledge Discovery and Dissemination Program, Creating an Intelligence Research Testbed, 2002–2003.

National Science Foundation (NSF), Information Technology Research Program COPLINK Center for Intelligence and Security Informatics Research, 2003–2005.

Department of Homeland Security (DHS), BorderSafe Program, Criminal Activity Network Analysis and Visualization, 2002–2005.

National Science Foundation (NSF), Disease Informatics Program, NV/BOT Portal: Developing a National Infectious Disease Information Infrastructure, 2003–2004.

References

Adderley, R., & Musgrove, P. B. (2001). Data mining case study: Modeling the behavior of offenders who commit serious sexual assaults. *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 215–220.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 207–216.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47–97.

Aleskerov, E., Freisleben, B., & Rao, B. (1997). CARDWATCH: A neural network based database mining system for credit card fraud detection. *Proceedings of Computational Intelligence for Financial Engineering (CIFEr)*, 220–226.

American Civil Liberties Union. (2004). *MATRIX: Myths and reality*. Retrieved July 27, 2004, from http://www.aclu.org/Privacy/Privacy.cfm?ID=14894&c=130

Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.

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

Arabie, P., Boorman, S. A., & Levitt, P. R. (1978). Constructing blockmodels: How and why. *Journal of Mathematical Psychology, 17*, 21–63.

Badiru, A. B., Karasz, J. M., & Holloway, B. T. (1988). AREST: Armed Robbery Eidetic Suspect Typing expert system. *Journal of Police Science and Administration, 16*, 210–216.

Baker, W. E., & Faulkner, R. R. (1993). The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry. *American Sociological Review, 58*(12), 837–860.

Baluja, S., Mittal, V., & Sukthankar, R. (1999). Applying machine learning for high performance named-entity extraction. In N. Cercone, K. Naruedomkul, & K. Kogure (Eds.), *PACLING '99: Proceedings of the Conference (Pacific Association of Computational Linguistics)* (pp. 1–14). Waterloo, Ont.: Department of Computer Science, University of Waterloo.

Bell, G. S., & Sethi, A. (2001). Matching records in a national medical patient index. *Communications of the ACM, 44*(9), 83–88.

Berndt, D. J., Bhat, S., Fisher, J. W., Hevner, A. R., & Studnicki, J. (2004). Data analytics for bioterrorism surveillance. *Proceedings of the Second Symposium on Intelligence and Security Informatics (ISI'04)*, 17–28.

Berndt, D. J., Hevner, A. R., & Studnicki, J. (2003). Bioterrorism surveillance with real-time data warehousing. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 322–335.

Borthwick, A., Sterling, J., Agichtein, E., & Grishman, R. (1998). NYU: Description of the MENE named entity system as used in MUC-7. *Proceedings of the 7th Message Understanding Conference (MUC-7)*. Retrieved August 19, 2004, from http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/nyu_english_named_entity.pdf

Bowen, J. E. (1994). An expert system for police investigators of economic crimes. *Expert Systems with Applications, 7*(2), 235–248.

Brahan, J. W., Lam, K. P., Chan, H., & Leung, W. (1998). AICAMS: Artificial Intelligence Crime Analysis and Management System. *Knowledge-Based Systems, 11*, 355–361.

Brantingham, P., & Brantingham, P. (1981). *Environmental criminology*. Beverly Hills, CA: Sage.

Brown, D. E. (1998). The Regional Crime Analysis Program (RECAP): A framework for mining data to catch criminals. *Proceedings of the 1998 International Conference on Systems, Man, and Cybernetics* (Vol. 3, pp. 2848–2853). Piscataway, NJ: IEEE.

Brown, D. E., Dalton, J., & Hoyle, H. (2004). Spatial forecast methods for terrorism events in urban environments. *Proceedings of the Second Symposium on Intelligence and Security Informatics (ISI'04)*, 426–435.

Brown, D. E., & Hagen, S. (2002). Data association methods with applications to law enforcement. *Decision Support Systems, 34*(4), 369–378.

Brown, D. E., & Oxford, R. B. (2001). Data mining time series with applications to crime analysis. *Proceedings of the 2001 IEEE International Conference on Systems, Man & Cybernetics Conference* (Vol. 3, pp. 1453–1458). Piscataway, NJ: IEEE.

Brown, M. (1998). *Future Alert Contact Network: Reducing crime via early notification.* Retrieved July 27, 2004, from http://pti.nw.dc.us/solutions/solutions98/public_safety/charlotte.html

Buccella, A., Cechich, A., & Brisaboa, N. R. (2003). An ontology approach to data integration. *Journal of Computer Science and Technology, 3*(2), 62–68.

Buetow, T., Chaboya, L., O'Toole, C., Cushna, T., Daspit, D., Peterson, T., et al. (2003). A spatial temporal visualizer for law enforcement. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 181–193.

Burt, R. S. (1976). Positions in networks. *Social Forces, 55*, 93–122.

Carley, K. M., Dombroski, M., Tsvetovat, M., Reminga, J., & Kamneva, N. (2003). Destabilizing dynamic covert networks. *Proceedings of the 8th International Command*

and Control Research and Technology Symposium. Retrieved August 19, 2004, from http://www.dodccrp.org/events/2003/8th_ICCRTS/pdf/021.pdf

Carley, K. M., Lee, J., & Krackhardt, D. (2002). Destabilizing networks. *Connections, 24*(3), 79–92.

Chan, P. K., & Stolfo, S. J. (1998). Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD'98),* 164–168.

Chau, M., Xu, J., & Chen, H. (2002). Extracting meaningful entities from police narrative reports. *Proceedings of the National Conference on Digital Government Research.* Retrieved August 19, 2004, from http://www.digitalgovernment.org/library/library/pdf/chau2.pdf

Chen, H. (2001). *Knowledge management systems: A text mining perspective.* Tucson: The University of Arizona.

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., Houston, A. L., Sewell, R. R., & Schatz, B. R. (1998). Internet browsing and searching: User evaluation of category map and concept space techniques. *Journal of the American Society for Information Science, 49*(7), 582–603.

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.

Chen, H., Miranda, R., Zeng, D. D., Demchak, C., Schroeder, J., & Madhusudan, T. (Eds.). (2003). *Intelligence and security informatics: Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics.* Berlin: Springer.

Chen, H., Moore, R., Zeng, D., & Leavitt, J. (Eds.). (2004). *Intelligence and security informatics: Proceedings of the Second Symposium on Intelligence and Security Informatics.* Berlin: Springer.

Chen, H., Qin, J., Reid, E., Chung, W., Zhou, Y., Xi, W., et al. (in press). The Dark Web Portal: Collecting and analyzing the presence of domestic and international terrorist groups on the Web. *Proceedings of the 7th Annual IEEE Conference on Intelligent Transportation Systems (ITSC 2004).*

Chen, H., Schroeder, J., Hauck, R., Ridgeway, L., Atabakhsh, H., Gupta, H., et al. (2003). COPLINK Connect: Information and knowledge management for law enforcement. *Decision Support Systems, 34*(3), 271–285.

Chen, H., Schuffels, C., & Orwig, R. (1996). Internet categorization and search: A self-organizing approach. *Journal of Visual Communication and Image Representation, 7*(1), 88–102.

Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W., & Schroeder, J. (2003). COPLINK: Managing law enforcement data and knowledge. *Communications of the ACM, 46*(1), 28–34.

Chen, I.-M. A., & Rotem, D. (1998). Integrating information from multiple independently developed data sources. *Proceedings of the 7th International Conference on Information and Knowledge Management,* 242–250.

Chinchor, N. A. (1998). Overview of MUC-7/MET-2. *Proceedings of the 7th Message Understanding Conference (MUC-7).* Retrieved August 19, 2004, from http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html

Coady, W. F. (1985). Automated link analysis: Artificial intelligence-based tool for investigators. *Police Chief, 52*(9), 22–23.

Collins, P. I., Johnson, G. F., Choy, A., Davidson, K. T., & Mackay, R. E. (1998). Advances in violent crime analysis and law enforcement: The Canadian Violent Crime Linkage Analysis System. *Journal of Government Information, 25*(3), 277–284.

Cook, J. S., & Cook, L. L. (2003). Social, ethical and legal issues of data mining. In J. Wang (Ed.), *Data mining: Opportunities and challenges* (pp. 395–420). Hershey, PA: Idea Group Publishing.

Craglia, M., Haining, R., & Wiles, P. (2000). A comparative evaluation of approaches to urban crime pattern analysis. *Urban Studies, 37*(4), 711–729.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. New York: Cambridge University Press.

Cronin, B. (2005). Intelligence, terrorism, and national security. *Annual Review of Information Science and Technology, 39*, 395–432.

Damianos, L., Ponte, J., Wohlever, S., Reeder, F., Day, D., Wilson, G., et al. (2002). MiTAP for bio-security: A case study. *AI Magazine, 23*(4), 13–29.

Davies, P. H. J. (2002). Intelligence, information technology, and information warfare. *Annual Review of Information Science and Technology, 36*, 313–352.

Defays, D. (1977). An efficient algorithm for a complete link method. *Computer Journal, 20*(4), 364–366.

Diederich, J., Kindermann, J., Leopold, E., & Paass, G. (2003). Authorship attribution with support vector machines. *Applied Intelligence, 19*(1–2), 109–123.

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

Dolotov, A., & Strickler, M. (2003). Web-Based Intelligence Reports System. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 39–58.

Duda, R. O., & Hart, P. E. (1973). *Pattern recognition and scene analysis*. New York: Wiley.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences, 95*(25), 14863–14868.

Eisenbeis, R., & Avery, R. (1972). *Discrimination analysis and classification procedures*. Lanham, MA: Lexington Books.

Estivill-Castro, V., & Lee, I. (2001). Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. *Proceedings of the 6th International Conference on GeoComputation*. Retrieved August 19, 2004, from http://www.geocomputation.org/2001/papers/estivillcastro.pdf

Evan, W. M. (1972). An organization-set model of interorganizational relations. In M. Tuite, R. Chisholm, & M. Radnor (Eds.), *Interorganizational decision-making* (pp. 181–200). Chicago: Aldine.

Faggiani, D., & McLaughlin, C. (1999). Using nation incident-based reporting system data for strategic crime analysis. *Journal of Quantitative Criminology, 15*(2), 181–191.

Fayyad, U. M., Djorgovshi, S. G., & Weir, N. (1996). Automating the analysis and cataloging of sky surveys. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 471–493). Menlo Park, CA: AAAI Press.

Fayyad, U., Piatetsk-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27–34.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks, 1*, 215–240.

Garcia-Molina, H., Ullman, J. D., & Widom, J. (2002). *Database systems: The complete book.* Upper Saddle River, NJ: Prentice-Hall.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24*, 189–199.

Gibson, D., Kleinberg, J., & Raghavan, P. (1998). Inferring Web communities from link topology. *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, 225–234.

Goldberg, D., Nichols, D., Oki, B., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM, 35*(12), 61–69.

Goldberg, H. G., & Senator, T. E. (1998). Restructuring databases for knowledge discovery by consolidation and link formation. *Proceedings of the 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis*, 47–52.

Goldberg, H. G., & Wong, R. W. H. (1998). Restructuring transactional data for link analysis in the FinCen AI System. *Proceedings of the 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis*, 38–46.

Grishman, R. (2003). Information extraction. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 545–559). New York: Oxford University Press.

Grubesic, T. H., & Murray, A. T. (2001). Detecting hot spots using cluster analysis and GIS. *Proceedings of 2001 Crime Mapping Research Conference.* Retrieved August 19, 2004, from http://www.ojp.usdoj.gov/nij/maps/Conferences/01conf/Grubesic.doc

Haas, L. M. (2002). Data integration through database federation. *IBM Systems Journal, 41*(4), 578–596.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques.* San Francisco, CA: Morgan Kaufmann.

Hand, D. J. (1981). *Discrimination and classification.* Chichester, UK: Wiley.

Harris, K. D. (1990). *Geographic factors in policing.* New York: McGraw-Hill.

Hasselbring, W. (2000). Information system integration. *Communications of the ACM, 43*(6), 33–38.

Hassibi, K. (2000). Detecting payment card fraud with neural networks. In P. J. G. Lisboa, A. Vellido, & B. Edisbury (Eds.), *Business applications of neural networks* (pp. 141–158). Singapore: World Scientific.

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. (Also available as Research Report No. MSR–TR–95–06 from Microsoft.)

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks, 13*, 415–425.

Icove, D. J. (1986). Automated crime profiling. *Law Enforcement Bulletin, 55*, 27–30.

Jain, A. K., & Flynn, P. J. (1996). Image segmentation using clustering. In N. Ahuja & K. Bowyer (Eds.), *Advances in image understanding* (pp. 65–83). Piscataway, NJ: IEEE Press.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Jhingran, A. D., Mattos, N., & Pirahesh, H. (2002). Information integration: A research agenda. *IBM Systems Journal, 41*(4), 555–562.

Kangas, L. J., Terrones, K. M., Keppel, R. D., & La Moria, R. D. (2003). Computer Aided Tracking and Characterization of Homicides and sexual assaults (CATCH). In J. Mena (Ed.), *Investigative data mining for security and criminal detection* (pp. 364–375). Amsterdam: Butterworth Heinemann.

Kasad, T., & Su, S. (2004). Transnational information sharing and event notification. In *Proceedings of the International Association for Development of the Information Society, International e-Society (IADIS e-Society '04)*, 52–62.

Kay, B. A., Timperi, R. J., Morse, S. S., Forslund, D., McGowan, J. J., & O'Brien, T. (1998). Innovative information-sharing strategies. *Emerging Infectious Diseases, 4*(3). Retrieved August 19, 2004, from http://www.cdc.gov/ncidod/eid/vol4no3/kay.htm

Klerks, P. (2001). The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. *Connections, 24*(3), 53–65.

Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer-Verlag.

Koperski, K., & Han, J. (1995). Discovery of spatial association rules in geographic information databases. *Proceedings of the 4th International Symposium on Large Spatial Databases (Advances in Spatial Databases)*, 47–66.

Krebs, V. E. (2001). Mapping networks of terrorist cells. *Connections, 24*(3), 43–52.

Krupka, G. R., & Hausman, K. (1998). IsoQuest Inc.: Description of the NetOwl text extractor system as used for MUC-7. *Proceedings of the 7th Message Understanding Conference (MUC-7)*. Retrieved August 19, 2004, from http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/isoquest.pdf

Kumar, A., & Olmeda, I. (1999). A study of composite or hybrid classifiers for knowledge discovery. *INFORMS Journal on Computing, 11*(3), 267–277.

Lee, R. (1998). Automatic information extraction from documents: A tool for intelligence and law enforcement analysts. *Proceedings of the 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis*, 63–67.

Lee, W., & Stolfo, S. (1998). Data mining approaches for intrusion detection. *Proceedings of the 7th USENIX Security Symposium*. Retrieved August 20, 2004, from http://citeseer.ist.psu.edu/cache/papers/cs/3327/http:zSzzSzwww.cs.columbia.eduzSz~wenkezSzpaperszSzusenix.pdf/lee98data.pdf

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

Levine, N. (2000). CrimeStat: A spatial statistics program for the analysis of crime incident locations. *Crime Mapping News, 2*(1), 8–9.

Lim, E.-P., Srivastava, J., Prabhakar, S., & Richardson, J. (1996). Entity identification in database integration. *Information Sciences, 89*, 1–38.

Lin, S., & Brown, D. E. (2003). Criminal incident data association using the OLAP technology. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 13–26.

Lippmann, R. P. (1987). An introduction to computing with neural networks. *IEEE Acoustics Speech and Signal Processing Magazine, 4*(2), 4–22.

Lorrain, F. P., & White, H. C. (1971). Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology, 1*, 49–80.

Lu, Q., Huang, Y., & Shekhar, S. (2003). Evacuation planning: A capacity constrained routing approach. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 111–125.

Mannila, H., Toivonen, H., & Inkeri, V. A. (1994). Efficient algorithms for discovering association rules. *Proceedings of Knowledge Discovery in Databases (KDD'94)*, 181–192.

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

McAndrew, D. (1999). The structural analysis of criminal networks. In D. Canter & L. Alison (Eds.), *The social psychology of crime: Groups, teams, and networks* (pp. 53–94). Dartmouth, UK: Aldershot.

McDonald, D., & Chen, H. (2002). Using sentence-selection heuristics to rank text segments in TXTRACTOR. *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'02)*, 28–35.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka, (Ed.), *Frontiers of econometrics* (pp. 105–142). New York: Academic Press.

McKeown, K., Barzilay, R., Chen, J., Elson, D., Evans, D., Klavans, J., et al. (2003). Columbia's Newsblaster: New features and future directions. *Proceedings of Human Language Technology Conference (HLT-NAACL 2003)*, 15–16.

Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., et al. (1998). BBN: Description of the SIFT system as used for MUC-7. *Proceedings of the 7th Message Understanding Conference (MUC-7)*. Retrieved August 19, 2004, from http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/bbn_muc7.pdf

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*, 17–23.

Murray, A. T., & Estivill-Castro, V. (1998). Cluster discovery techniques for exploratory spatial data analysis. *International Journal of Geographical Information Science, 12*, 431–443.

Murray, A. T., McGuffog, I., Western, J. S., & Mullins, P. (2001). Exploratory spatial data analysis techniques for examining urban crime. *British Journal of Criminology, 41*, 309–329.

National Research Council. (2002). *Making the nation safer: The role of science and technology in countering terrorism*. Washington, DC: National Academy Press.

Newcombe, H. B., Kennedy, J. M., Axford S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science, 130*(3381), 954–959.

O'Hara, C. E., & O'Hara, G. L. (1980). *Fundamentals of criminal investigation* (5th ed.). Springfield, IL: Charles C. Thomas.

Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis, 27*, 286–296.

Patman, F., & Thompson, P. (2003). Names: A new frontier in text mining. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 27–38.

Pinner, R. W., Rebmann, C. A., Schuchat, A., & Hughes, J. M. (2003). Disease surveillance and the academic, clinical, and public health communities. *Emerging Infectious Diseases, 9*(7). Retrieved August 19, 2004, from http://www.cdc.gov/ncidod/eid/vol9no7/03-0083.htm

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

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*: San Francisco: Morgan Kaufmann.

Raghu, T. S., Ramesh, R., & Whinston, A. B. (2003). Addressing the homeland security problem: A collaborative decision-making framework. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 249–265.

Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal, 10*, 334–350.

Rasmussen, E. (1992). Clustering algorithms. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms* (pp. 419–442). Englewood Cliffs, NJ: Prentice Hall.

Ratchliffe, J. H., & McCullagh, M. J. (1999). Hotbeds of crime and the search for spatial accuracy. *Journal of Geographical Systems, 1*(4), 385–398.

Reid, E., Qin, J., Chung, W., Xu, J., Zhou, Y., Schumaker, R., et al. (2004). Terrorism Knowledge Discovery Project: A knowledge discovery approach to address the threats of terrorism. *Proceedings of the Second Symposium on Intelligence and Security Informatics (ISI'04)*, 125–145.

Riloff, E. (1996). Automatically generating extraction patterns from untagged text. *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI '96)*, 1044–1049.

Ronfeldt, D., & Arquilla, J. (2001). What next for networks and netwars? In J. Arquilla & D. Ronfeldt (Eds.), *Networks and netwars: The future of terror, crime, and militancy* (pp. 311–362). Santa Monica, CA: Rand Press.

Rossmo, D. K. (1995). Overview: Multivariate spatial profiles as a tool in crime investigation. In C. R. Block, M. Dabdoub, & S. Fregly (Eds.), *Crime analysis through computer mapping* (pp. 65–97). Washington, DC: Police Executive Research Forum.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McLelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 318–362). Cambridge, MA: MIT Press.

Ryan, J., Lin, M., & Mikkulainen, R. (1998). Intrusion detection with neural networks. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems* (pp. 943–949). Cambridge, MA: MIT Press.

Saether, M., & Canter, D. V. (2001). A structural analysis of fraud and armed robbery networks in Norway. *Proceedings of the 6th International Investigative Psychology Conference*. Retrieved August 20, 2004, from http://www.i-psy.com/conferences/sixth/multimedia/powerPoint/saether/saether.htm

Sageman, M. (2004). *Understanding terror networks*. Philadelphia: University of Pennsylvania Press.

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

Schroeder, J., Xu, J., & Chen, H. (2003). CrimeLink Explorer: Using domain knowledge to facilitate automated crime association analysis. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 168–180.

Schumacher, B. J., & Leitner, M. (1999). Spatial crime displacement resulting from large-scale urban renewal programs in the city of Baltimore, MD: A GIS modeling approach.

Proceedings of the 4th International Conference on GeoComputation. Retrieved August 19, 2004, from http://www.geocomputation.org/1999/047/gc_047.htm

Shortliffe, E. H., & Blois, M. S. (2000). The computer meets medicine and biology: Emergence of a discipline. In K. J. Hannah & M. J. Ball (Eds.), *Health informatics* (pp. 1–40). New York: Springer-Verlag.

Silverstein, C., Brin, S., & Motwani, R. (1998). Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery, 2*, 39–68.

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

Stolfo, S. J., Hershkop, S., Wang, K., Nimeskern, O., & Hu, C.-W. (2003). Behavior profiling and email. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 74–90.

Strickland, L. S. (2002a). Information and the war against terrorism (Part I). *Bulletin of the American Society for Information Science and Technology, 28*(2), 12–17.

Strickland, L. S. (2002b). Information and the war against terrorism (Part II): Were American intelligence and law enforcement effectively positioned to protect the public? *Bulletin of the American Society for Information Science and Technology, 28*(3), 18–22.

Strickland, L. S. (2002c). Information and the war against terrorism (Part III): New information-related laws and the impact on civil liberties. *Bulletin of the American Society for Information Science and Technology, 29*(3), 23–27.

Strickland, L. S. (2002d). Information and the war against terrorism (Part IV): Civil liberties vs. security in the age of terrorism. *Bulletin of the American Society for Information Science and Technology, 28*(4), 9–13.

Strickland, L. S. (2002e). Information and the war against terrorism (Part V): The business implications. *Bulletin of the American Society for Information Science and Technology, 28*(6), 18–21.

Strickland, L. S. (2005). Domestic security surveillance and civil liberties. *Annual Review of Information Science and Technology, 39*, 433–513.

Sun, A., Naing, M.-M., Lim, E.-P., & Lam, W. (2003). Using support vector machines for terrorism information extraction. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 1–12.

Tolle, K. M., & Chen, H. (2000). Comparing noun phrasing techniques for use with medical digital library tools. *Journal of the American Society for Information Science, 51*(4), 352–370.

Torgerson, W. S. (1952). Multidimensional scaling: Theory and method. *Psychometrika, 17*, 401–419.

Trybula, W. J. (1999). Text mining. *Annual Review of Information Science and Technology, 34*, 385–419.

Tufte, E. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

U.S. Federal Bureau of Investigation. (1992). *Uniform crime reporting handbook: National incident-based reporting system (NIBRS)*. Washington, DC: The Bureau.

U.S. Office of Homeland Security. (2002). *National Strategy for Homeland Security*. Washington DC: Office of Homeland Security. Retrieved August 19, 2004, from http://www.whitehouse.gov/homeland/book/nat_strat_hls.pdf

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

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

Wang, J.-H., Huang, C.-C., Teng, J.-W., & Chien, L.-F. (2004). Generating concept hierarchies from text for intelligence analysis. *Proceedings of the Second Symposium on Intelligence and Security Informatics (ISI'04)*, 100–113.

Wang, J.-H., Lin, B. T., Shieh, C.-C., & Deng, P. S. (2003). Criminal record matching based on the vector space model. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 386.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* Cambridge, UK: Cambridge University Press.

Weisburd, D., & McEwen, T. (Eds.). (1997). *Crime mapping and crime prevention.* Monsey, NY: Criminal Justice Press.

Weiss, S. I., & Kulikowski, C. A. (1991). *Computer systems that learn: Classification and prediction methods from statistics, neural networks, machine learning, and expert systems.* San Francisco: Morgan Kaufmann.

Williams, C. (1975). Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. *Biometrika, 62*, 207–212.

Witten, I. H., Bray, Z., Mahoui, M., & Teahan, W. J. (1999). Using language models for generic entity extraction. *Proceedings of the ICML Workshop on Text Mining.* Retrieved August 20, 2004, from http://www-ai.ijs.si/DunjaMladenic/ICML99/WittenFinal.ps

Xu, J., & Chen, H. (2003). Untangling criminal networks: A case study. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 232–248.

Xu, J., & Chen, H. (in press). Criminal network analysis and visualization: A data mining perspective. *Communications of the ACM.*

Xue, Y., & Brown, D. E. (2003). Decision based spatial analysis of crime. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 153–167.

Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B. T., & Liu, X. (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems, 14*(4), 32–43.

Zeng, D., Chen, H., Daspit, D., Shan, F., Nandiraju, S., Chau, M., et al. (2003). COPLINK Agent: An architecture for information monitoring and sharing in law enforcement. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 281–295.

Zeng, D., Chen, H., Tseng, C., Larson, C., Eidson, M., Gotham, I., et al. (2004). West Nile virus and botulism portal: A case study in infectious disease informatics. *Proceedings of the Second Symposium on Intelligence and Security Informatics (ISI'04)*, 28–41.

Zhang, Z., Salerno, J. J., & Yu, P. S. (2003). Applying data mining in investigating money laundering crimes. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 747–752.

Zhao, J. L., Bi, H. H., & Chen, H. (2003). Collaborative workflow management for interagency crime analysis. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 266–280.

Zheng, R., Qin, Y., Huang, Z., & Chen, H. (2003). Authorship analysis in cybercrime investigation. *Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI'03)*, 59–73.