Chapter 3

ISI RESEARCH: LITERATURE REVIEW

Chapter Overview

In this chapter, we review the technical foundations of ISI and the six classes of data mining technologies specified in our ISI research framework: information sharing and collaboration, crime association mining, crime classification and clustering, intelligence text mining, spatial and temporal crime pattern mining, and criminal network analysis. We also summarize relevant research that addresses knowledge discovery in public safety and national security.

3.1 Information Sharing and Collaboration

Information sharing across jurisdictional boundaries of intelligence and security agencies has been identified as one of the key foundations for securing national security (Office of Homeland Security, 2002). However, sharing and integrating information from distributed, heterogeneous, and autonomous data sources is a non-trivial task (Hasselbring, 2000; Rahm and 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 and Rotem, 1998).

Three approaches to data integration have been proposed: *federation*, *warehousing*, and *mediation* (Garcia-Molina et al., 2002). Database federation maintains data in their original, independent sources but provides a uniform data access mechanism (Buccella et al., 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. Many hybrid approaches have been proposed (Jhingran et al., 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 et al., 1996; Rahm and Bernstein, 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 and 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 et al., 1959) to match attributes from different databases (Bell and 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 database. 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, or (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 et al., 2003b) employed the database federation approach to achieve schema-level data integration. It provided a common COPLINK schema and a one-stop-shopping user interface to facilitate the access of different data sources from multiple police departments. Evaluation results showed that COPLINK Connect had improved performance over the Record Management System (RMS) of police data in system effectiveness, ease of use, and interface design (Chen et al., 2003b). Similarly, the Phoenix Police Department Reports (PPDR) is a web-based, federated intelligence system in which databases shared common schema (Dolotov and 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 et al., 2004; Berndt et al., 2003).

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

Information sharing has also been undertaken in intelligence and security agencies through cross-jurisdictional collaborative systems. COPLINK Agent was run on top of the COPLINK Connect system (Chen et al., 2003b) and linked crime investigators who were working on related crime cases at different units to enhance collaboration (Zeng et al., 2003). It employed collaborative filtering approaches (Goldberg et al., 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 (Brown, 1998b). Its collaboration capability, however, was relatively limited. Research has also been performed to mathematically model collaboration processes across law enforcement and intelligence jurisdictions to improve work productivity (Raghu et al., 2003; Zhao et al., 2003). Although information sharing and collaboration are not knowledge discovery *per se*, they prepare data for important subsequent analyses.

3.2 Crime Association Mining

Finding associations among data items is an important topic in knowledge discovery research. One of the most widely studied approaches is association rule mining, a process of 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 what other products (Agrawal et al., 1993; Mannila et al., 1994; Silverstein et al., 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, "diapers $\Rightarrow$ milk with 60% *support* and 90% *confidence*" means that 60% of customers buy both diapers and milk in the same transaction and that 90% of the customers who buy diapers tend to also buy milk.

In the intelligence and security domain, spatial association rule mining (Koperski and Han, 1995) has been proposed to extract cause-effect relations among geographically referenced crime data to identify environmental factors that attract crimes (Estivill-Castro and 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 and Brown, 2003).

The purpose of incident association mining is to find crimes that might be committed by the same offender so that unsolved crimes can be linked to solved crimes to identify the suspect. This technique is often used to solve serial crimes such as serial sexual offenses and serial homicide. However, finding associated crime incidents can be fairly 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 1 million human hours (Brown and Hagen, 2002). When the number of crime incidents is large, manual identification of associations between crimes is prohibitive in

time. Two approaches, *similarity-based* and *outlier-based*, have been developed for incident association mining.

Similarity-based methods detect associations between crime incidents by comparing crimes' features such as spatial locations of the incidents and offenders' modus operandi (MO), which often are regarded as an offender's "behavioral signature" (O'Hara and 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 et al., 1998). However, as the collection of human-decision rules requires a large amount of 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 (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 such as an offender's eye color that takes on categorical values, they developed a similarity table that specified the similarity level for each pair of categorical values based on heuristics. Evaluation showed that this approach enhanced both accuracy and efficiency for associating crime records. Similarly, Wang et al. (Wang et al., 2003) proposed to measure 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 spaces. The vector space model is widely employed in information retrieval applications and various similarity measures could be used (Rasmussen, 1992).

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 and Brown, 2003). For example, in a series of robberies a Japanese sword was used as the weapon. Since a Japanese sword is a very uncommon weapon as opposed to other weapons such as shotguns, it is very likely that this series of robberies was committed by the same offender. Based on this outlier concept, crime investigators need to first cluster crime 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 showed that the outlier-based method was 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 often is referred to as entity association mining (Lin and Brown, 2003) or link analysis (Sparrow, 1991) in law enforcement. The purpose is to find out how crime entities that appear

to be unrelated at the surface are actually linked to each other. Law enforcement officers and crime investigators throughout the world have long used link analysis to search for and analyze relationships between 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).

Three types of link analysis approaches have been suggested: *heuristic-based*, *statistical-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 (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 have 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 and Senator, 1998; Goldberg and Wong, 1998). The COPLINK Detect system (Hauck et al., 2002) employed a statistical-based approach called Concept Space (Chen and 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 is. Zhang et al. (Zhang et al., 2003) proposed to use a fuzzy resemblance function to calculate the correlation between two individuals' past financial transactions 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 primarily used to identify associations between entities extracted from textual documents such as police report narratives. Lee (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 (Coady, 1985) proposed to use the PROLOG language to automatically derive rules of entity associations from text data and use the rules to detect associations in similar documents. Template-based approaches heavily rely on a fixed set of predefined patterns and rules, and thus may have a limited scope of application.

3.3 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 and Kulikowski, 1991). Examples of classification applications include fraud detection (Chan and Stolfo, 1998), computer and network intrusion detection (Lee and Stolfo, 1998), bank failure prediction (Sarkar and Sriram, 2001), and image categorization (Fayyad et al., 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. Accuracy of the classifier is assessed using the testing set.

Discriminant analysis (Eisenbeis and Avery, 1972), Bayesian models (Duda and Hart, 1973; Heckerman, 1995), decision trees (Quinlan, 1986, 1993), artificial neural networks (Rumelhart et al., 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 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 the data items, and an output layer, with nodes representing the class membership labels. Neural networks learn and encode knowledge through the connection weights.

SVM is a novel learning classifier based on the Structural Risk Minimization principle from the computational learning theory. SVM is capable of handling millions of inputs and does not require feature selection (Cristianini and 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 and 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, to identify fraudulent financial transactions Aleskerov et al. (Aleskerov et al., 1997) employed neural networks to detect anomalies in customers' credit card transactions based on their transaction history. Hassibi (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 and Stolfo, 1998). Misuse detection identifies attacks by matching them to previously known attack patterns or signatures. Anomaly detection, on the other hand, identifies abnormal user behaviors based on historical data. Lee and Stolfo (Lee and Stolfo, 1998) employed decision rule induction approaches to classify *sendmail* system call traces into normal or abnormal traces. Ryan et al. (Ryan et al., 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 et al. (Stolfo et al., 2003) applied Bayesian classifiers to distinguish between normal email and spamming email.

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 inter-cluster similarity (Jain et al., 1999). Clustering has been used in a variety of applications, including image segmentation (Jain and Flynn, 1996), gene clustering (Eisen et al., 1998), and document categorization (Chen et al., 1998; Chen et al., 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. 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 the problem of determining the appropriate number of clusters (Jain et al., 1999). Different from 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 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 to find a geographical area with a high concentration of certain types of crimes. For example, Adderley and Musgrove (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) 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 et al. (Kangas et al., 2003) employed the SOM method to group crime incidents to identify serial murderers and sexual offenders. Brown (Brown, 1998a) proposed *k*-means and the nearest neighbor approach to clustering spatial data of crimes to find "hotspot" areas in a city. The spatial clustering methods are often used in "hotspot analysis," which will be reviewed in detail in Section 3.5.

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 intensiveness and frequency of relationships between offenders. Stolfo et al. (Stolfo et al., 2003) proposed to group email users that frequently communicate with each other into clusters so that unusual email behavior that violated the group communication patterns could be effectively detected. Offender clustering is more often used in criminal network analysis, which will be reviewed in detail in Section 3.6.

3.4 Intelligence Text Mining

A large amount of intelligence- and security-related data is represented in text forms such as police narrative reports, court transcripts, news clips, and web articles. Valuable information in such texts is often difficult to retrieve, access, and use for the purposes of crime investigation and counter-terrorism. It is desirable to automatically mine the text data 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 automatically identify from text documents the names of entities of interest, such as persons (e.g., "John Doe"), locations (e.g., "Washington, D.C."), 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, email addresses, and web addresses (URLs). The Message Understanding Conference (MUC) series has been the major forum for researchers in this area, where researchers meet and compare the performance of their entity extraction approaches (Chinchor, 1998).

Four major named-entity extraction approaches have been proposed: lexical-lookup, rule-based, statistical model, 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 organization names in the U.S., all person last names obtained from the government census data, etc. These systems work by looking up phrases in texts that match the items specified in their lexicons (Borthwick et al., 1998).
- *Rule-based*. Rule-based systems rely on hand-crafted rules to identify named entities. The rules may be structural, contextual, or lexical (Krupka and Hausman, 1998). An example rule would look like the following:
  capitalized last name + , + capitalized first name $\Rightarrow$ *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 model*. Such systems often use statistical models to identify occurrences of certain cues or particular patterns for entities in texts. A training dataset is needed for a system to obtain the statistics. The statistical language model reported in (Witten et al., 1999) is an example for such systems.
- *Machine learning*. This type of system relies on machine learning algorithms rather than human-created rules to extract knowledge or identify patterns from text data. Examples of machine learning algorithms used in entity extraction include neural networks, decision trees (Baluja et al., 1999), Hidden Markov Model (Miller et al., 1998), and entropy maximization (Borthwick et al., 1998).

Instead of relying on a single approach, most existing information extraction systems utilize a combination of two or more of these approaches. Many systems were evaluated at the MUC-7 Conference. The best systems were able to achieve over 90% in both precision and recall rates in extracting persons, locations, organizations, dates, times, currencies, and percentages.

Recent years have seen research on named-entity extraction for intelligence and security applications (Patman and Thompson, 2003; Wang et al., 2004b). For example, Chau et al. (Chau et al., 2002) developed a

neural network-based entity extraction system to identify person 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% precision and recall rates for person name and narcotic drug names. However, it was difficult to achieve satisfactory performance for addresses and personal properties because of their wide coverage. Sun et al. (Sun et al., 2003) converted the entity extraction problem into a classification problem to identify relevant entities from the MUC text collection in the terrorism domain. They first identified all noun phrases in a document and then used the support vector machine to classify these entity candidates based on both content and context features. The results showed that for the specific terrorism text collection the approach's performance in precision and *F* measure was comparable with AutoSlog (Riloff, 1996), one of the best entity extraction systems.

Several news and event extraction systems have been reported recently, including Columbia's Newsblaster (McKeown et al., 2003) and Carnegie Mellon University's system (Yang et al., 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.

3.5 Crime Spatial and Temporal Mining

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

Spatial pattern analysis and geographical profiling of crimes play important roles in solving crimes (Rosmo, 1995). Three approaches for crime spatial pattern mining have been reported: *visual approach*, *clustering approaches*, and *statistical approaches* (Murray et al., 2001). The visual approach is also called crime mapping. It presents a city or region 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 the relationship between spatial features and crime occurrences.

The clustering approach has been used in hotspot analysis, a process of automatically identifying areas with high crime concentration. This type of analysis helps law enforcement effectively allocate police efforts to reduce crimes in hotspot areas. Partitional clustering algorithms such as the *k*-means methods are often used for finding crime hotspots (Murray and Estivill-Castro, 1998). For example, Schumacher and Leitner (Schumacher and Leitner, 1999) used the *k*-means algorithm to identify hotspots in the downtown areas of Baltimore. Comparing the hotspots of different years, they found spatial patterns of the displacement of crimes after redevelopment of the downtown area. Corresponding proactive strategies were then suggested based on the patterns found. Although they are efficient and scalable compared with hierarchical clustering algorithms, partitional clustering algorithms usually require the user to predefine the number of clusters to be found. This is not always feasible, however (Grubesic and Murray, 2001). Accordingly, researchers have tried to use statistical approaches to conduct hotspot analysis or to test the significance of hotspots (Craglia et al., 2000). The test statistics G_i (Getis and Ord, 1992; Ord and Getis, 1995) and Moran's I (Moran, 1950), which are used to test the significance of spatial autocorrelation, can be used to detect hotspots. If a variable is correlated with itself through space it is said to be spatially autocorrelated. For example, Ratchliffe and McCullagh (Ratchliffe and McCullagh, 1999) employed the G_i and G_i^* statistics to identify the hotspots of residential burglary and motor vehicle crimes in a city. Compared with a domain expert's perception of the hotspots, this approach was shown to be effective.

Statistical approaches have also been used in crime prediction applications. Based on the spatial choice theory (McFadden, 1973), Xue and Brown (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 hotspot models. Similarly, Brown et al. (Brown et al., 2004) built a logistic regression model to predict suicide bombing in counter-terrorism applications.

Commercially available graphical 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 and 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 the *statistical approach*. Visualization approaches present individual or aggregated temporal features of crimes using periodic view or timeline view. Common methods for viewing periodic data include sequence charts, point charts, bar charts, line charts, and spiral graphs displayed in 2D or 3D (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 et al., 2003) seamlessly incorporated periodic view, timeline view, and GIS view in the system to provide support to crime 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 (Brown and Oxford, 2001) developed several statistical models, including log-normal regression model, Poisson regression model, and cumulative logistic regression model, to predict the number of breaking and entering crimes. The log-normal regression model was found to fit the data best (Brown and Oxford, 2001).

3.6 Criminal Network Analysis

Criminals seldom operate alone 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 effective operation of a criminal enterprise. Criminal enterprises can be viewed as a network consisting of nodes (individual offenders) and links (relationships). In criminal networks, groups or teams may exist within which members have close relationships. One group also may interact with other groups to obtain or transfer illicit goods. Moreover, individuals play different roles in their groups. For example, some key members may act as leaders to control the activities of a group. Some others may serve as gatekeepers to ensure the smooth flow of information or illicit goods.

Structural network patterns in terms of subgroups, between-group interactions, and individual roles thus are important to 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 crimes. For example, removal of central members in a network

may effectively upset the operational network and put a criminal enterprise out of action (Baker and 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 and 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.

Social Network Analysis (SNA) provides a set of measures and approaches for structural network analysis (Wasserman and Faust, 1994). These techniques were originally designed to discover social structures in social networks (Wasserman and 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 and Faulkner, 1993; Saether and Canter, 2001). In general, SNA is capable of detecting subgroups, identifying central individuals, discovering between-group interaction patterns, and uncovering a network's 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 and Faust, 1994). Burt (Burt, 1976) proposed to apply hierarchical clustering methods based on a structural equivalence measure (Lorrain and White, 1971) to partition a social network into positions in which members have similar structural roles. Xu and Chen (Xu and Chen, 2003) employed hierarchical clustering to detect criminal groups in narcotics networks 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 (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 interpret 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 and Faulkner, 1993). Krebs found that, in the network consisting of the nineteen September 11th hijackers, Mohamed Atta had the highest degree score (Krebs, 2001).

- *Discovery of patterns of interaction*. Patterns of interaction between subgroups can be discovered using an SNA approach called blockmodel analysis (Arabie et al., 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 and 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 they are (McAndrew, 1999). Several network analysis tools, such as Analyst's Notebook (Klerks, 2001), Netmap (Goldberg and Senator, 1998), and Watson (Anderson et al., 1994), can automatically draw a graphical representation of a criminal network. However, they do not provide much structural analysis functionality and rely on investigators' manual examinations to extract structural patterns.

The above-reviewed six classes of KDD techniques 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 et al., 2003; Carley et al., 2002).

3.7 Future Directions

There are many opportunities for researchers from different disciplines to contribute to the science of ISI. Computer science researchers in databases, artificial intelligence, data mining, algorithms, networking, and grid computing are poised to contribute significantly to core information infrastructure, integration, and analysis research of relevance to ISI. Information systems, information science, and management science

researchers could help develop the quantitative, system, and information theory-based methodologies needed for the systematic study of national security. Cognitive science, behavioral research, management and policy, and legal disciplines are critical to the understanding of the individual, group, organizational, and societal impacts and the creation of effective national security policies. All the abovementioned academic disciplines need to work closely with the domain experts in law enforcement, public safety, emergency response, and the intelligence community.

3.8 Questions for Discussion

1. What are the core technical foundations for ISI research? What are the undergraduate and graduate courses in various disciplines that can help contribute to this understanding?
2. What are the major conferences, journals, and magazines that publish and promote ISI research? Are there professional ISI-related societies or work groups?
3. What are the major research centers, laboratories, non-profit agencies, and government programs that provide training, education, and research resources of relevance to ISI?
4. Are there government sponsored national security-related training and fellowship programs? What about some of the intern, co-op, and employment opportunities for an IT career in national security?