Crime Data Mining: A General Framework and Some Examples

By increasing efficiency and reducing errors, crime data mining techniques can facilitate police work and enable investigators to allocate their time to other valuable tasks.

Hsinchun Chen
Wingyan Chung
Jennifer Jie Xu
Gang Wang
Yi Qin
University of Arizona

Michael Chau
University of Hong Kong

Concern about national security has increased significantly since the terrorist attacks on 11 September 2001. The CIA, FBI, and other federal agencies are actively collecting domestic and foreign intelligence to prevent future attacks. These efforts have in turn motivated local authorities to more closely monitor criminal activities in their own jurisdictions.

A major challenge facing all law-enforcement and intelligence-gathering organizations is accurately and efficiently analyzing the growing volumes of crime data. For example, complex conspiracies are often difficult to unravel because information on suspects can be geographically diffuse and span long periods of time. Detecting cybercrime can likewise be difficult because busy network traffic and frequent online transactions generate large amounts of data, only a small portion of which relates to illegal activities.

Data mining is a powerful tool that enables criminal investigators who may lack extensive training as data analysts to explore large databases quickly and efficiently.[1] Computers can process thousands of instructions in seconds, saving precious time. In addition, installing and running software often costs less than hiring and training personnel. Computers are also less prone to errors than human investigators, especially those who work long hours.

We present a general framework for crime data mining that draws on experience gained with the Coplink project (http://ai.bpa.arizona.edu/coplink), which researchers at the University of Arizona have been conducting in collaboration with the Tucson and Phoenix police departments since 1997.

CRIME TYPES AND SECURITY CONCERNS

A criminal act can encompass a wide range of activities, from civil infractions such as illegal parking to internationally organized mass murder such as the 9/11 attacks. Law-enforcement agencies across the US compile crime statistics using well-established standards such as the FBI's Uniform Crime Reporting System and its successor, the National Incident-Based Reporting System (www. fbi.gov/hq/cjisd/ucr.htm), as well as other criteria defined by jurisdictional needs and requirements.

Table 1 lists eight crime categories on which local and federal authorities maintain data, ordered by their increasing degree of harm to the general public. We devised these categories, which include numerous offenses classified by different law-enforcement agencies in various ways, in consultation with a local detective with more than 30 years of experience.

Some types of crime, such as traffic violations and arson, primarily concern police at the city, county, and state levels. Other crime types are investigated by local law-enforcement units as well as by national and international agencies. For example, a city police department's sex crimes unit may track local pedophiles and prostitutes, while the FBI and the International Criminal Police Organization focus on transnational trafficking in children and women for sexual exploitation.

Many crimes, such as the theft of nuclear weapons data, can have profound implications for both

Published by the IEEE Computer Society
0018-9162/04/$20.00 © 2004 IEEE

Table 1. Crime types at different law-enforcement levels.

Crime type	Local law enforcement	National and international security
Traffic violations	Speeding, reckless driving, causing property damage or personal injury in a collision, driving under the influence of drugs or alcohol, hit-and-run, "road rage"	—
Sex crime	Sexual abuse, rape, sexual assault, child molestation, child pornography, prostitution	Trafficking in women and children for sexual exploitation, including prostitution and pornography
Theft	Robbery, burglary, larceny, motor vehicle theft	Theft of national secrets or weapon information, illicit trafficking in stolen art and vehicles
Fraud	Money laundering, counterfeiting, insurance fraud, corruption and bribery, misappropriation of assets	Transnational money laundering, fraud, and corruption; trafficking in stolen software, music, movies, and other intellectual property
Arson	Intentionally setting fires to damage property, such as a warehouse or apartment building	—
Gang/drug offenses	Possessing, distributing, and selling illegal drugs	Transnational drug trafficking, organized racketeering and extortion, people smuggling
Violent crime	Murder, aggravated assault, armed robbery, forcible rape, hate crime	Terrorism, air and maritime piracy, bombings
Cybercrime	Internet fraud, such as credit card and advance fee fraud, fraudulent Web sites, and illegal online gambling and trading; network intrusion and hacking; virus spreading; cyberpiracy and cyberterrorism; distributing child pornography; identity theft	

national and global security. Transnational fraud and trafficking in stolen property or contraband can severely impact trade, business, and government revenue. Local gangs as well as foreign-based drug cartels and criminal organizations exact a large financial cost as well as threaten public health and safety. Although most types of violent crime—such as murder, robbery, forcible rape, and aggravated assault—are local police matters, terrorism is a global problem that relies on cooperation at all levels of government. The Internet's pervasiveness likewise makes identity theft, network intrusion, cyberpiracy, and other illicit computer-mediated activities a challenge for many law-enforcement bodies.[2]

CRIME DATA MINING TECHNIQUES

Traditional data mining techniques such as association analysis, classification and prediction, cluster analysis, and outlier analysis identify patterns in structured data.[3] Newer techniques identify patterns from both structured and unstructured data. As with other forms of data mining, crime data mining raises privacy concerns.[4] Nevertheless, researchers have developed various automated data mining techniques for both local law enforcement and national security applications.

Entity extraction identifies particular patterns from data such as text, images, or audio materials. It has been used to automatically identify persons, addresses, vehicles, and personal characteristics from police narrative reports.[5] In computer forensics, the extraction of software metrics[6]—which includes the data structure, program flow, organization and quantity of comments, and use of variable names—can facilitate further investigation by, for example, grouping similar programs written by hackers and tracing their behavior. Entity extraction provides basic information for crime analysis, but its performance depends greatly on the availability of extensive amounts of clean input data.

Clustering techniques group data items into classes with similar characteristics to maximize or minimize intraclass similarity—for example, to identify suspects who conduct crimes in similar ways or distinguish among groups belonging to different gangs. These techniques do not have a set of predefined classes for assigning items. Some researchers use the statistics-based *concept space* algorithm to automatically associate different objects such as persons, organizations, and vehicles in crime records.[7] Using link analysis techniques to identify similar transactions, the Financial Crimes Enforcement Network AI System[8] exploits Bank Secrecy Act data to support the detection and analysis of money laundering and other financial crimes. Clustering crime incidents can automate a major part of crime analysis but is limited by the high computational intensity typically required.

Association rule mining discovers frequently occurring item sets in a database and presents the patterns as rules. This technique has been applied in network intrusion detection to derive association rules from users' interaction history. Investigators also can apply this technique to network intruders' profiles to help detect potential future network attacks.[9]

Similar to association rule mining, *sequential pattern mining* finds frequently occurring sequences of items over a set of transactions that occurred at different times. In network intrusion detection, this approach can identify intrusion patterns among

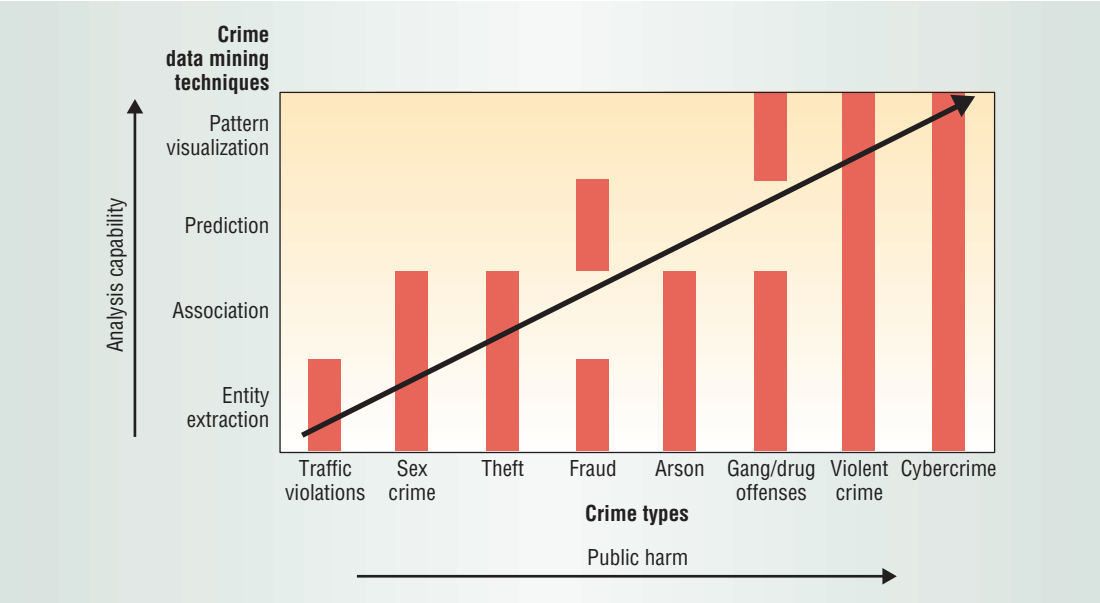

Figure 1. Crime data mining framework. The framework identifies relationships between techniques applied in criminal and intelligence analysis at the local, national, and international levels.

time-stamped data. Showing hidden patterns benefits crime analysis, but to obtain meaningful results requires rich and highly structured data.

Deviation detection uses specific measures to study data that differs markedly from the rest of the data. Also called *outlier detection*, investigators can apply this technique to fraud detection, network intrusion detection, and other crime analyses. However, such activities can sometimes appear to be normal, making it difficult to identify outliers.

Classification finds common properties among different crime entities and organizes them into predefined classes. This technique has been used to identify the source of e-mail spamming based on the sender's linguistic patterns and structural features.[10] Often used to predict crime trends, classification can reduce the time required to identify crime entities. However, the technique requires a predefined classification scheme. Classification also requires reasonably complete training and testing data because a high degree of missing data would limit prediction accuracy.

String comparator techniques compare the textual fields in pairs of database records and compute the similarity between the records. These techniques can detect deceptive information—such as name, address, and Social Security number—in criminal records.[11] Investigators can use string comparators to analyze textual data, but the techniques often require intensive computation.

Social network analysis describes the roles of and interactions among nodes in a conceptual network. Investigators can use this technique to construct a network that illustrates criminals' roles, the flow of tangible and intangible goods and information, and associations among these entities. Further analysis can reveal critical roles and subgroups and vulnerabilities inside the network. This approach enables visualization of criminal networks, but investigators still might not be able to discover the network's true leaders if they keep a low profile.

CRIME DATA MINING FRAMEWORK

Many efforts have used automated techniques to analyze different types of crimes, but without a unifying framework describing how to apply them. In particular, understanding the relationship between analysis capability and crime type characteristics can help investigators more effectively use those techniques to identify trends and patterns, address problem areas, and even predict crimes.

Based on the Tucson Police Department's crime classification database, which contains approximately 1.3 million suspect and criminal records ranging from 1970 to the present, and on the existing literature, we have developed the general framework for crime data mining shown in Figure 1.

The framework shows relationships between data mining techniques applied in criminal and intelligence analysis and the crime types listed in Table 1. The vertical axis arranges the techniques in increasing order of analysis capability. We identified four major categories of crime data mining techniques: entity extraction, association, prediction, and pattern visualization. Each category represents a set of techniques for use in certain types of crime analysis.

For example, investigators can use neural network techniques in crime entity extraction and prediction. Clustering techniques are effective in crime association and prediction. Social network analysis can facilitate crime association and pattern visualization. Investigators can apply various techniques independently or jointly to tackle particular crime analysis problems.

Guided by our detective consultant, we have arranged the eight crime types in increasing order of public harm on the horizontal axis. The shaded regions represent research using various analytical

techniques on certain crime types. Although they can apply any technique to any crime type, when highly organized criminal enterprises involve many people and have a pervasive societal impact, investigators must apply a spectrum of techniques to discover associations, identify patterns, and make predictions.

We believe that our framework has general applicability to crime and intelligence analysis because it encompasses all major crime types as well as both traditional and new intelligence-specific data mining techniques.

COPLINK CASE STUDY

To illustrate our crime data mining framework, we describe three examples of its use in the Coplink project: named-entity extraction, deceptive-identity detection, and criminal-network analysis.

Named-entity extraction

Most criminal justice databases capture only structured data that fits in predefined fields. Our first data mining task involved extracting named entities from police narrative reports, which are difficult to analyze using automated techniques. We randomly selected 36 narcotics-related reports from the Phoenix Police Department that were relatively *noisy*—all were written in uppercase letters and contained many typos, spelling errors, and grammatical mistakes.

We adopted a modified version of the AI Entity Extractor system, which uses a three-step process to identify the names of persons, locations, and organizations in a document. First, it identifies noun phrases according to linguistic rules. Second, the system calculates a set of feature scores for each phrase based on pattern matching and lexical lookup. Third, it uses a feedforward/backpropagation neural network to predict the most likely entity type for each phrase.

The AI Entity Extractor has been compared to systems reported at the Sixth Message Understanding Conference (MUC-6) and achieved above-average performance. To adopt the system for crime analysis applications, we modified it to identify five entity types: person names, addresses, vehicles, narcotic names, and physical characteristics.[5]

Working from the selected reports, our detective consultant manually identified all entities that belong to the five categories of interest. We then conducted threefold cross-validation testing to evaluate the system.

The modified extractor performed well in identifying person names (74.1 percent) and narcotic drugs (85.4 percent) from the test data set, but not as well for addresses (59.6 percent) and personal properties (46.8 percent). Recalls for the same categories were 73.4, 77.9, 51.4, and 47.8 percent, respectively. Vehicle name results were not analyzed because only four references to vehicles occurred in the 36 reports.

These preliminary results demonstrated the feasibility and potential value of applying entity extraction techniques to crime data mining, especially considering that the narrative reports were much noisier than the news articles used in the MUC-6 evaluations. Although we tested only 36 reports in our study, we plan to use training and testing data typical of other practical data mining applications to more thoroughly evaluate the system. We also are exploring interactive ways to integrate human knowledge into the extractor's learning component.

Using advanced analysis techniques to further extract information from narrative documents, such as the roles of entities and the relationships among them, is an important future research direction. Often, entities provide the unit of analysis in other crime data mining applications.

Crime investigators must apply a spectrum of techniques to discover associations, identify patterns, and make predictions.

Deceptive-identity detection

Suspects often give false names, birth dates, or addresses to police officers and thus have multiple database entries, making it difficult for officers to determine a suspect's true identity and relate past incidents involving that person. Our second data mining task involved automatically detecting deceptive criminal identities from the Tucson Police Department's database, which contains information such as name, gender, address, ID number, and physical description. Our detective consultant manually identified 120 deceptive criminal records involving 44 suspects from the database.

Based on the criminal identity deception taxonomy we developed in a case study, we selected name, birth date, address, and Social Security number to represent a criminal's identity and ignored other less reliable fields. Our method employed string comparators to compare values in the corresponding fields of each record pair.[11] Comparators measure the similarity between two strings. We normalized the similarity values between 0 and 1, and calculated an overall similarity measure between two records as a Euclidean vector norm over the four chosen fields. A Euclidean vector norm is the square root of the sum of squared similarity measures and is also normalized between 0 and 1.

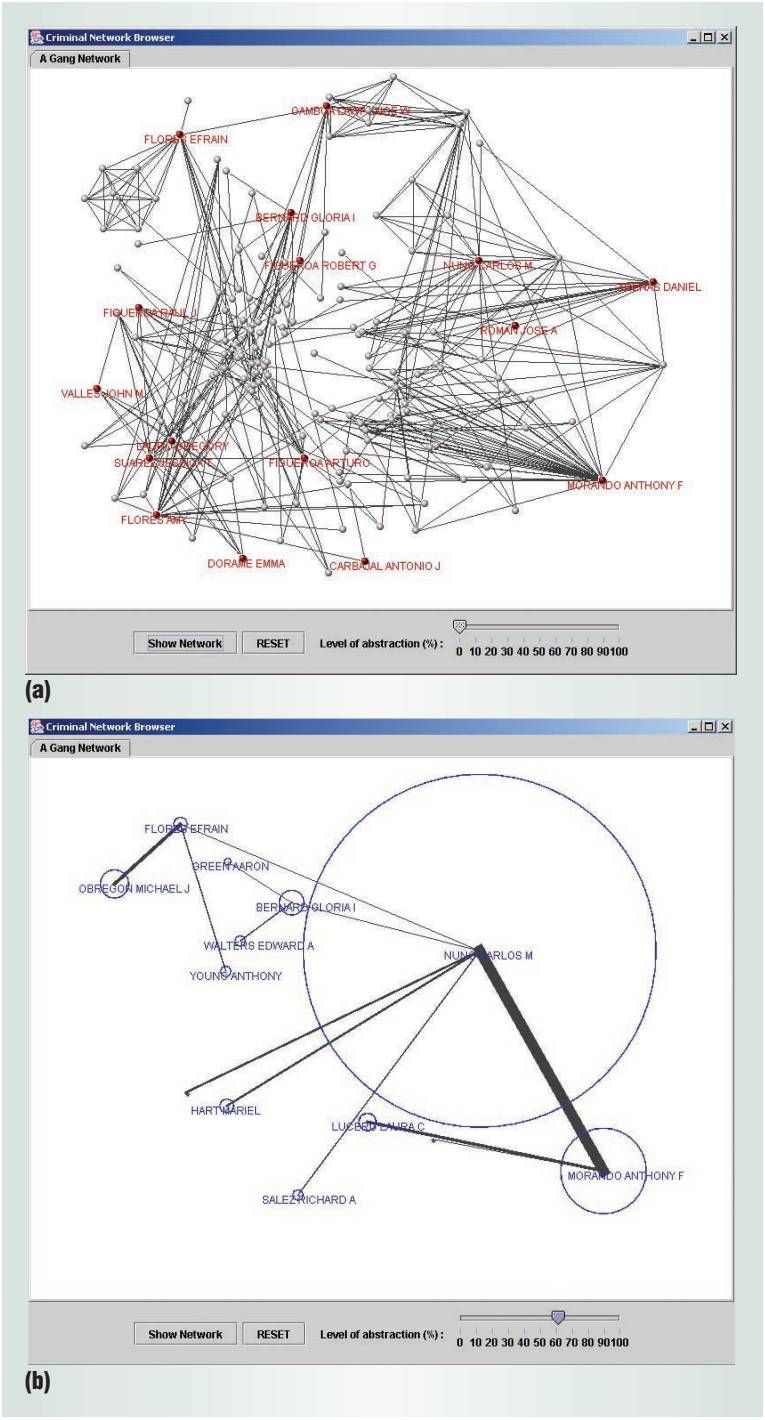

Figure 2. Criminal-network analysis. (a) Data mining uncovered 16 target gang members from a network of 164 criminals in the Tucson Police Department database. (b) The circles represent subgroups and are tagged with their leaders' names. Each circle's size is proportional to the number of members in the subgroup.

We employed a holdout validation method using two-thirds of the data for training and the rest for testing. In the training stage, we tried threshold values ranging from 0.00 to 1.00 that differentiated between deceptive and not-deceptive records. We first determined the optimal threshold to be reached when the association decisions best matched our expert's judgments, 0.48, then used this value to assess our algorithm's predicted association accuracy in the testing stage.

In the training stage, the algorithm achieved its highest accuracy, 97.4 percent, with a low 2.6 percent false-negative rate and 2.6 percent false-positive rate. In the testing stage, the algorithm achieved an accuracy of 94.0 percent in linking deceptive records that pointed to the same suspect.

The results again demonstrated that crime data mining is feasible and promising. Testing errors that occurred in the false-negative category, in which unrelated suspects were recognized as being related, might be caused by the overall threshold value obtained from the training stage. Thus, an adaptive threshold might be more desirable for developing an automated process in future research.

With this technique, law-enforcement officers can retrieve existing identity records relating to a suspect in their databases that traditional exact-match techniques often fail to locate. They also can use this technique as a preprocessing tool to combine identity records representing the same person for other crime data mining applications, thereby improving subsequent data analysis.

Criminal-network analysis

Criminals often develop networks in which they form groups or teams to carry out various illegal activities. Our third data mining task consisted of identifying subgroups and key members in such networks and then studying interaction patterns to develop effective strategies for disrupting the networks. Our data came from 272 Tucson Police Department incident summaries involving 164 crimes committed from 1985 through May 2002.

We used a concept-space approach to extract criminal relations from the incident summaries and create a likely network of suspects. Co-occurrence weight measured the relational strength between two criminals by computing how frequently they were identified in the same incident.[7] We used hierarchical clustering to partition the network into subgroups and the block-modeling approach to identify interaction patterns between these subgroups.[12] We also calculated centrality measures—degree, betweenness, and closeness—to detect key members in each group, such as leaders and gatekeepers.

As Figure 2a shows, data mining uncovered 16 target gang members from the resulting network. In Figure 2b, the circles represent subgroups the system found, and they bear the labels of their leaders' names. A circle's size is proportional to the number of members in that subgroup. The thick-

ness of straight lines connecting circles indicates the strength of relationships between subgroups.

We conducted a two-hour field study with three Tucson Police Department domain experts who evaluated the analysis's validity by comparing the results against their knowledge of gang organization. They confirmed that the system-found subgroups correctly represented the real groups' organization. For example, the biggest group consisted of gang members involved in many murders and assaults. The second largest group specialized in drug distribution and sales. Interaction patterns between subgroups found in the network were valid as well.

According to the experts, members from the two biggest groups associated frequently, and their leaders were good friends. In most cases, the analysis also correctly identified central members who played important roles. For example, the leader of the second largest subgroup made considerable money from selling and distributing drugs.

All three experts believed that this system could greatly increase crime analysts' work productivity by efficiently extracting criminal association information from data and using that information to generate and visualize criminal networks. More importantly, it would help discover knowledge about criminal organizations that requires many hours to uncover manually. In addition, the system would help novice investigators understand the structure and operations of criminal enterprises relatively quickly. Finally, it could suggest investigative leads that would otherwise be overlooked and help prevent crimes by disrupting criminal networks effectively.

Studying criminal networks requires additional data mining capabilities: entity extraction and co-occurrence analysis to identify criminal entities and associations, clustering and block modeling for discovering subgroups and interaction patterns, and visualization for presenting analysis results. One drawback of our current approach is that it generates mostly static networks. Given that criminal networks are dynamic, future research will focus on the evolution and prediction of criminal networks.

Human investigators with years of experience can often analyze crime trends precisely, but as the incidence and complexity of crime increases, human errors occur, analysis time increases, and criminals have more time to destroy evidence and escape arrest. By increasing efficiency and reducing errors, crime data mining techniques can facilitate police work and enable investigators to allocate their time to other valuable tasks.

Much work remains in this emerging field. For example, investigators can use crime entity-extraction techniques to analyze the behavioral patterns of serial offenders. Crime association and clustering techniques can reveal the identities of cybercriminals who use the Internet to spread illegal messages or malicious code. Investigators can use machine-learning algorithms—such as ID3, neural networks, Support Vector Machines, and genetic algorithms—to predict crimes by analyzing factors such as time, location, vehicle, address, physical characteristics, and property. They also can use these tools to develop more intuitive techniques for crime pattern and network visualization.

We are continuing our research in crime data mining using Coplink data as our testbed. We intend to supplement this work with information from the US Secret Service and the US Citizenship and Immigration Services, public infrastructure data, terrorists' and extremists' Web sites, television news archives, and disease and biological agent statistics. Currently, we are creating a multilingual cybercrime database with the help of experts from the Tucson and Phoenix police departments and the Taiwan Criminal Investigation Bureau, our research partner. This effort will monitor suspicious Internet newsgroups, chat rooms, peer-to-peer network messages, and Web sites in the US and Taiwan and download messages for further analysis. ■

Acknowledgment

The Coplink project was funded by the National Institute of Justice and the National Science Foundation.

References

1. U.M. Fayyad and R. Uthurusamy, "Evolving Data Mining into Solutions for Insights," *Comm. ACM*, Aug. 2002, pp. 28-31.
2. W. Chang et al., "An International Perspective on Fighting Cybercrime," *Proc. 1st NSF/NIJ Symp. Intelligence and Security Informatics*, LNCS 2665, Springer-Verlag, 2003, pp. 379-384.
3. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001.
4. H. Kargupta, K. Liu, and J. Ryan, "Privacy-Sensitive Distributed Data Mining from Multi-Party Data," *Proc. 1st NSF/NIJ Symp. Intelligence and Security Informatics*, LNCS 2665, Springer-Verlag, 2003, pp. 336-342.

5. M.Chau, J.J. Xu, and H. Chen, "Extracting Meaningful Entities from Police Narrative Reports, *Proc. Nat'l Conf. Digital Government Research*, Digital Government Research Center, 2002, pp. 271-275.
6. A. Gray, P. Sallis, and S. MacDonell, "Software Forensics: Extending Authorship Analysis Techniques to Computer Programs," *Proc. 3rd Biannual Conf. Int'l Assoc. Forensic Linguistics*, Int'l Assoc. Forensic Linguistics, 1997, pp. 1-8.
7. R.V. Hauck et al., "Using Coplink to Analyze Criminal-Justice Data," *Computer*, Mar. 2002, pp. 30-37.
8. T. Senator et al., "The FinCEN Artificial Intelligence System: Identifying Potential Money Laundering from Reports of Large Cash Transactions," *AI Magazine*, vol.16, no. 4, 1995, pp. 21-39.
9. W. Lee, S.J. Stolfo, and W. Mok, "A Data Mining Framework for Building Intrusion Detection Models," *Proc. 1999 IEEE Symp. Security and Privacy*, IEEE CS Press, 1999, pp. 120-132.
10. O. de Vel et al., "Mining E-Mail Content for Author Identification Forensics," *SIGMOD Record*, vol. 30, no. 4, 2001, pp. 55-64.
11. G. Wang, H. Chen, and H. Atabakhsh, "Automatically Detecting Deceptive Criminal Identities," *Comm. ACM*, Mar. 2004, pp. 70-76.
12. S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge Univ. Press, 1994.

Hsinchun Chen *is the McClelland Professor of Management Information Systems at the University of Arizona's Eller College of Business and Administration. His research interests include medical informatics, homeland security, semantic retrieval, search algorithms, knowledge management, and Web computing. Chen received a PhD in information systems from New York University. Contact him at hchen@eller.arizona.edu.*

Wingyan Chung *is a doctoral candidate in the Department of Management Information Systems at the University of Arizona, where he is a research associate in the Artificial Intelligence Lab. His research interests include knowledge management, knowledge discovery on the Web, text mining, security informatics, and human-computer interaction. Chung received an MS in information and technology management from the Chinese University of Hong Kong. He is a member of the IEEE Computer Society, the ACM, and the Association for Information Systems. Contact him at wchung@eller.arizona.edu.*

Jennifer Jie Xu *is a doctoral candidate in the Department of Management Information Systems at the University of Arizona, where she is a member of the Artificial Intelligence Lab. Her research interests include knowledge management, social network analysis, computer-mediated communication, and information visualization. Xu received an MS in computer science from the University of Mississippi. She is a member of the IEEE Computer Society. Contact her at jxu@eller.arizona.edu.*

Gang Wang *is a doctoral student in the Department of Management Information Systems at the University of Arizona. His research interests include deception detection, data mining, Web mining, and knowledge discovery. Wang received an MS in industrial engineering from Louisiana State University. Contact him at gang@eller.arizona.edu.*

Yi Qin *was a member of the Artificial Intelligence Lab at the University of Arizona. Her research interests include medical informatics and cybercrime analysis. Qin received an MS in management information systems from the University of Arizona. Contact her at yiqin@eller.arizona.edu.*

Michael Chau *is a research assistant professor in the School of Business at the University of Hong Kong and was formerly a research associate at the University of Arizona's Artificial Intelligence Lab. His research interests include text mining, Web mining, digital libraries, knowledge management, and intelligent agents. Chau received a PhD in management information systems from the University of Arizona. He is a member of the IEEE Computer Society, the ACM, the Association for Information Systems, and the American Society for Information Science and Technology. Contact him at mchau@business.hku.hk.*