

# Crime Data Mining: An Overview and Case Studies

Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng,  
Homa Atabakhsh

{hchen, wchung, yiqin, mchau, jxu, gang, rong, homa}@bpa.arizona.edu

Artificial Intelligence Lab, Department of Management Information Systems, University of Arizona,  
Tucson, AZ 85721, USA

<http://ai.bpa.arizona.edu/>

**Abstract.** The concern about national security has increased significantly since the 9/11 attacks. However, information overload hinders the effective analysis of criminal and terrorist activities. Data mining applied in the context of law enforcement and intelligence analysis holds the promise of alleviating such problems. In this paper, we review crime data mining techniques and present four case studies done in our ongoing COPLINK project.

## 1. Introduction

The concern about national security has increased significantly since the terrorist attacks on September 11, 2001. Intelligence agencies such as the CIA and FBI are actively collecting and analyzing information to investigate terrorists' activities. Local law enforcement agencies have also become more alert to criminal activities in their own jurisdictions. One challenge to law enforcement and intelligence agencies is the difficulty of analyzing large volumes of data involved in criminal and terrorist activities. Data mining holds the promise of making it easy, convenient, and practical to explore very large databases for organizations and users. In this paper, we review data mining techniques applied in the context of law enforcement and intelligence analysis, and present four case studies done in our ongoing COPLINK project (Hauck et al., 2002).

## 2. An Overview of Crime Data Mining

It is useful to review crime data mining in two dimensions: (1) crime types and security concerns and (2) crime data mining approaches and techniques.

### 2.1 Crime Types and Security Concerns

*Crime* is defined as “an act or the commission of an act that is forbidden, or the omission of a duty that is commanded by a public law and that makes the offender liable to punishment by that law” (Webster Dictionary). An act of crime encompasses a wide range of activities, ranging from simple violation of civic duties (e.g., illegal parking) to internationally organized crimes (e.g., the 9/11 attacks). Table 1 summarizes the different types of crimes in increasing degree of public influence. Note that both local and national law enforcement and security agencies are facing many similar challenges.

### 2.2 Crime Data Mining Approaches and Techniques

Data mining is defined as the identification of interesting structure in data, where structure designates patterns, statistical or predictive models of the data, and relationships among parts of the data (Fayyad & Uthurusamy, 2002). Data mining in the context of crime and intelligence analysis for national security is still a young field. The following describes our applications of different techniques in crime data mining. *Entity extraction* has been used to automatically identify person, address, vehicle, narcotic drug, and personal properties from police narrative reports (Chau et al., 2002). *Clustering techniques* such as “concept space” have been used to automatically associate different objects (such as persons, organizations, vehicles) in crime records (Hauck et al., 2002). *Deviation detection* has been applied in fraud detection, network intrusion detection, and other crime analyses that involve tracing abnormal activities. *Classification* has been used to detect email spamming and find authors who send out unsolicited emails (de Vel et al., 2001). *String comparator* has been used to detect deceptive information in criminal records (Wang et al., 2002). *Social network analysis* has been used to analyze criminals' roles and associations among entities in a criminal network.

Increasing public influence ↓	Type	Local Law Enforcement Level	National Security Level
	Traffic Violations	Driving under influence (DUI), fatal/personal injury/property damage traffic accident, road rage	-
	Sex Crime	Sexual offenses, sexual assaults, child molesting	Organized prostitution
	Theft	Robbery, burglary, larceny, motor vehicle theft, stolen property	Theft of national secrets or weapon information
	Fraud	Forgery and counterfeiting, frauds, embezzlement, identity deception	Transnational money laundering, identity fraud, transnational financial fraud
	Arson	Arson on buildings, apartments	-
	Gang / drug offenses	Narcotic drug offenses (sales or possession)	Transnational drug trafficking
	Violent Crime	Criminal homicide, armed robbery, aggravated assault, other assaults	Terrorism (bioterrorism, bombing, hijacking, etc.)
	Cyber Crime	Internet frauds, illegal trading, network intrusion/hacking, virus spreading, hate crimes, cyber-piracy, cyber-pornography, cyber-terrorism, theft of confidential information	

**Table 1. Crime types at different levels**

### 3. Case Studies of Crime Data Mining

Based on the crime characteristics and analysis techniques discussed above, we present four case studies of crime data mining that are part of our ongoing COPLINK project.

#### 3.1 Entity Extraction for Police Narrative Reports

Valuable criminal-justice data in free texts such as police narrative reports are currently difficult to be accessed and used by intelligence investigators in crime analyses. We proposed a neural network-based entity extractor, which applies named-entity extraction techniques to automatically identify useful entities from police narrative reports of the Tucson Police Department (TPD). The system has three major components: (1) *Noun phrasing*: It is a modified version of the Arizona Noun Phraser (Tolle & Chen, 2000) and extracts noun phrases as named entities from documents based on syntactical analysis; (2) *Finite state machine and lexical lookup*: A finite state machine compares each word in the extracted phrase, as well as the words immediately before and after the phrase, with the items in a handcrafted lexicons. Each comparison will generate a binary value (either 0 or 1) to indicate a match or mismatch; (3) *Neural network*: The feedforward/backpropagation neural network predicts the phrase's most possible entity type. Preliminary evaluation results demonstrated that our approach is feasible and has some potential values for real-life applications. Our system achieved encouraging precision and recall rates for person names and narcotic drugs (74 – 85%), but did not perform as well for addresses and personal properties (47 – 60%) (Chau et al., 2002). Our future work includes conducting larger-scale evaluation studies and enhancing the system to capture human knowledge interactively.

#### 3.2 Detecting Criminal Identity Deceptions: An Algorithmic Approach

Criminals often provide police officers with deceptive identities to mislead police investigations, for example, using aliases, fabricated birth dates or addresses, etc. The large amount of data also prevents officers from examining inexact matches manually. Based on a case study on deceptive criminal identities recorded in the TPD, we have built a taxonomy of criminal identity deceptions that consisted of name deceptions, address deceptions, date-of-birth deceptions, and identity number deceptions. We found criminals usually made minor changes to their real identity information. For example, one may give a name similarly spelled or, change the sequence of digits in his social security number. Based on the taxonomy, we developed an algorithmic approach to detect deceptive criminal identities automatically (Wang et al., 2002). Our approach utilized four identity fields: name, address, date-of-birth, and social-security-number and compared each corresponding field for a pair of criminal identity records. An overall disagreement value between the two records was computed by calculating the Euclidean Distance of disagreement measures over all attribute fields. A deception in this record pair will be noticed when the

overall disagreement value exceeds a pre-determined threshold value, which is acquired during training processes. We conducted an experiment using a sample set of real criminal identity records from the TPD. The results showed that our algorithm could accurately detect 94% of criminal identity deceptions.

### 3.3 Authorship Analysis in Cybercrime

The large amount of cyber space activities and their anonymous nature make cybercrime investigation extremely difficult. Conventional ways to deal with this problem rely on a manual effort, which is largely limited by the sheer amount of messages and constantly changing author IDs. We have proposed an authorship analysis framework to automatically trace identities of cyber criminals through messages they post on the Internet. Under this framework, three types of message features, including style markers, structural features, and content-specific features, are extracted and inductive learning algorithms are used to build feature-based models to identify authorship of illegal messages. To evaluate the effectiveness of this framework, we conducted an experimental study on data sets of English and Chinese email and online newsgroup messages produced by a small number of authors. We tested three inductive learning algorithms: decision trees, backpropagation neural networks, and Support Vector Machines. Our experiments demonstrated that with a set of carefully selected features and an effective learning algorithm, we were able to identify the authors of Internet newsgroup and email messages with a reasonably high accuracy. We achieved average prediction accuracies of 80% - 90% for email messages, 90% - 97% for the newsgroup messages, and 70% - 85% for Chinese Bulletin Board System (BBS) messages. Significant performance improvement was observed when structural features were added on top of style markers. SVM outperformed the other two classifiers on all occasions. The experimental results indicated a promising future of using our framework to address the identity-tracing problem.

### 3.4 Criminal Network Analysis

In organized crimes such as narcotics trafficking, terrorism, gang-related crimes, and frauds, offenders often cooperate and form networks to carry out various illegal activities. Social Network Analysis (SNA) has been recognized as an appropriate methodology to uncover previously unknown structural patterns from criminal networks. We have employed SNA techniques for criminal network analysis in our COPLINK project. Four steps are involved in our analysis: (1) *Network extraction*: We utilized TPD crime incident reports as sources for criminal relationship information because criminals who committed crimes together usually were related. The concept space approach (Chen & Lynch, 1992) was used to identify and uncover criminal relationships; (2) *Subgroup detection*: We employed hierarchical clustering to detect subgroups in a criminal network based on relational strength; (3) *Interaction pattern discovery*: We employed an SNA approach called blockmodeling to reveal patterns of between-group interaction. Given a partitioned network, blockmodel analysis determines the presence or absence of an interaction between a pair of subgroups by comparing the density of the links between these two subgroups to a predefined threshold value; (4) *Central member identification*: We employed several measures, such as degree, betweenness, and closeness to identify central members in a given subgroup. These three measures can suggest the centrality of a network member.

Figure 1 shows a narcotics network consisting of 60 criminals. It is difficult to detect subgroups, interaction patterns, and the overall structure from this original network manually. Using clustering and blockmodeling methods, however, a chain structure became apparent (Figure 1b). We have conducted a field study at the TPD with three domain experts who confirmed that the subgroups and central members found by the system were correct representations of the reality. They believed that this system would be very useful for crime investigation and could greatly increase crime analysts' productivity.

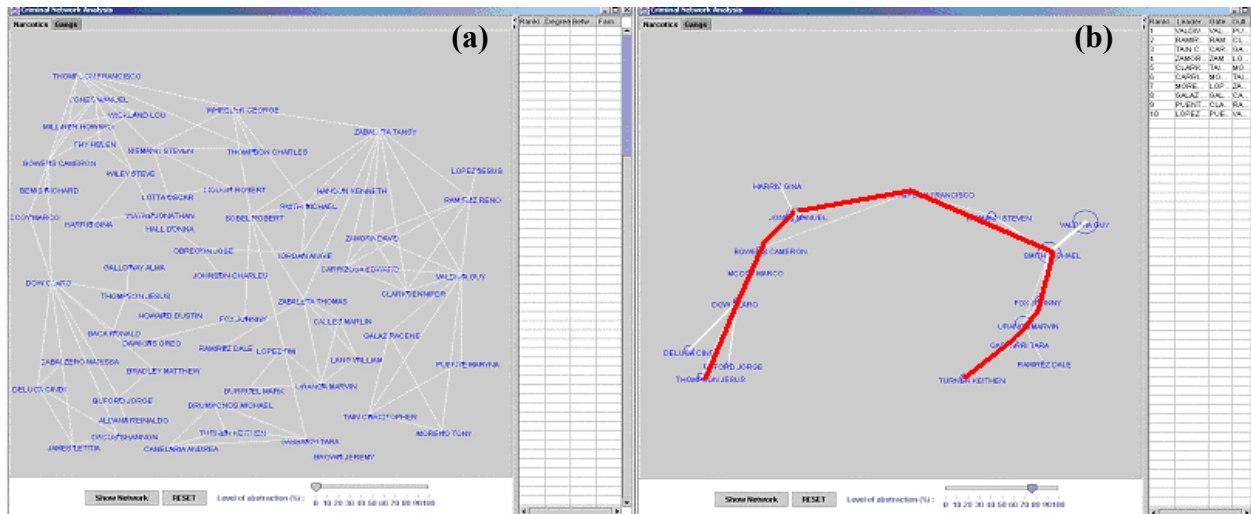


Figure 1(a). A 60-member narcotics network. (b) The chain structure (thicker links) found.

#### 4. Conclusions and Future Directions

In this paper, we have presented an overview of crime data mining and four COPLINK case studies. From the encouraging results, we believe that crime data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis. Many future directions can be explored in this still young field. For example, more visual and intuitive criminal and intelligence investigation techniques can be developed for crime pattern and network visualization.

#### 5. Acknowledgements

This project has primarily been funded by NSF Digital Government Program, “COPLINK Center: Information and Knowledge Management for Law Enforcement,” #9983304, July 2000 – June 2003. We would like to thank the following people for their support and assistance during the entire project development and evaluation process: All members of the University of Arizona Artificial Intelligence Lab staff and specifically past and present COPLINK team members; Lt. Jenny Schroeder, Detective Tim Petersen and other contributing personnel from the Tucson Police Department; and other contributing members of the Phoenix Police Department.

#### 6. References

Chau, M., Xu, J., & Chen, H. (2002). Extracting meaningful entities from police narrative reports. In: Proceedings of the National Conference for Digital Government Research (dg.o 2002), Los Angeles, California, USA.

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.

de Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining E-mail Content for Author Identification Forensics. *SIGMOD Record*, 30(4), 55-64.

Fayyad, U.M., & Uthurusamy, R. (2002). Evolving data mining into solutions for insights. *Communications of the ACM*, 45(8), 28-31.

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.

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.

Wang, G., Chen, H., & Atabakhsh, H. Automatically detecting deceptive criminal identities. *Communications of the ACM* (Accepted for publication, forthcoming).