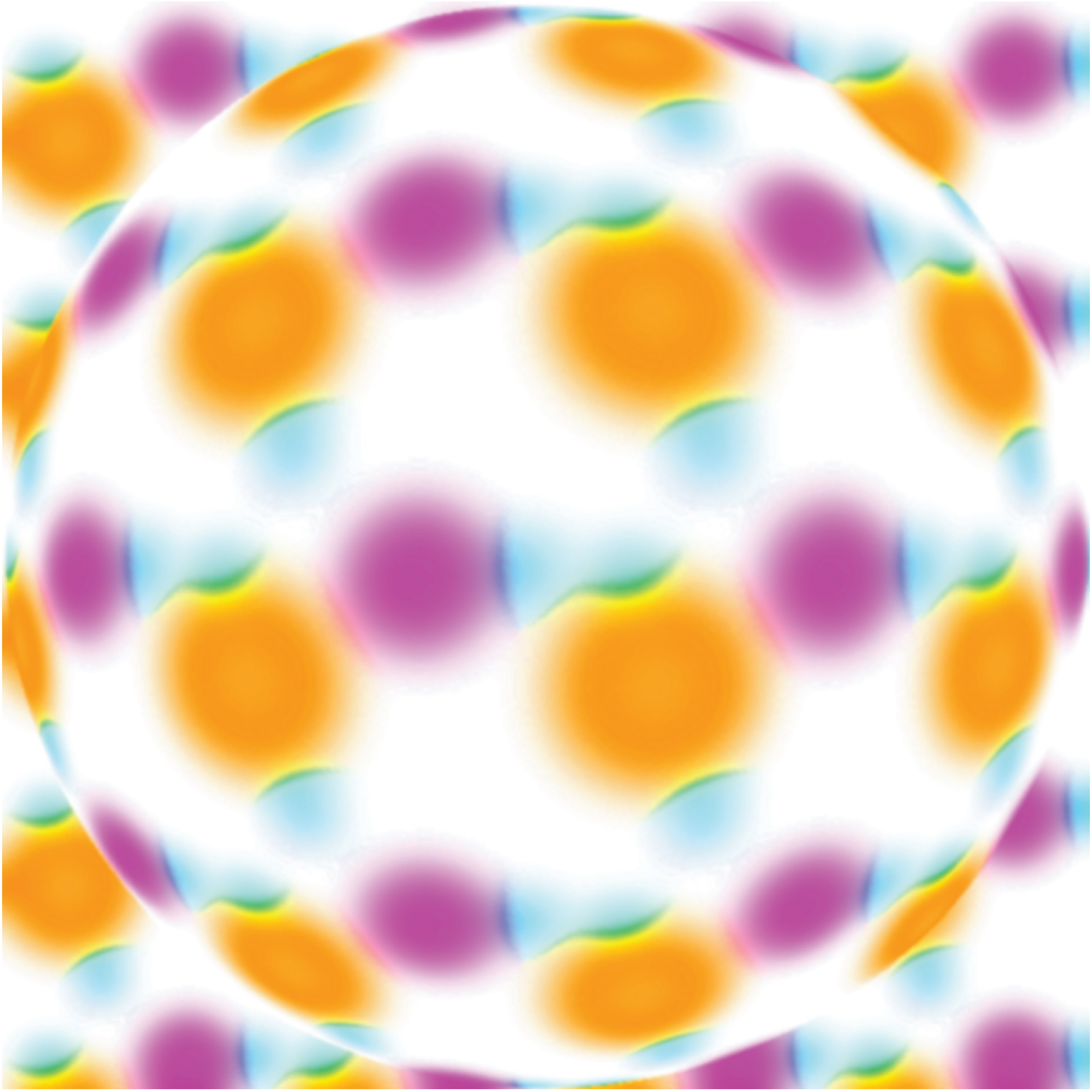


By Jennifer Xu and Hsinchun Chen

Criminal Network Analysis and Visualization

A new generation of data mining tools and applications work to unearth hidden patterns in large volumes of crime data.

A great deal has been written over the last four years regarding how academics have contributed possible technological solutions for uncovering terrorist networks to enhance public safety and national security. Both the public and the Pentagon have realized that knowledge of the structure of terrorist networks and how those networks operate is one of the key factors in winning the so-called "netwar." Probably the most critical weapons our intelligence and law enforcement agencies should carry are reliable data and sophisticated techniques that help discover useful knowledge from the data.

The study of terrorist networks falls into the larger category of criminal network analysis, which is often applied to investigations of organized crimes (for example, terrorism, narcotics trafficking, fraud, gang-related crimes, armed robbery, and so on). Unlike other types of crimes often committed by single or a few offenders, organized crimes are carried out by multiple, collaborating offenders, who may form groups and teams and play different roles. In a narcotics network, for instance, different groups may be responsible for handling the drug supply, distribution, sales, smuggling, and money laundering. In each group, there may be a leader who issues commands and provides steering mechanisms to the group, as well as gatekeepers who ensure that information and drugs flow effectively to and from other groups. Criminal network analysis therefore requires the ability to integrate information from multiple crime incidents or even multiple sources and discover regular patterns about the structure, organization, operation, and information flow in criminal networks.

To untangle and disrupt criminal networks, both reliable data and sophisticated techniques are indispensable. However, intelligence and law enforcement agencies are often faced with the dilemma of having too much data, which in effect makes too little value. On one hand, they have large volumes of "raw data" collected from multiple sources: phone records, bank accounts and transactions, vehicle sales and registration records, and surveillance reports, to name a few [9, 10]. On the other hand, they lack sophisticated network analysis tools and techniques to utilize the data effectively and efficiently.

Today's criminal network analysis is primarily a manual process that consumes much human time and efforts, thus has limited applicability to crime investigation. Our objective here is to provide a data mining perspective for criminal network analysis. We discuss the challenges in data processing, review existing network analysis and visualization tools, and recommend the Social Network Analysis (SNA) approaches. Although SNA is not traditionally considered as a data mining technique, it is especially suitable for mining large volumes of association data to discover hidden structural patterns in criminal networks [9, 10]. We also report some data mining projects for criminal network analysis in the COPLINK research, which is the NIJ- and NSF- funded research for management of law enforcement knowledge [3].

Challenges in Data Processing

Like data mining applications in many other domains, mining law enforcement data involves many obstacles. First, incomplete, incorrect, or inconsistent data can create problems. Moreover, these characteristics of criminal networks cause difficulties not common in other data mining applications:

- *Incompleteness.* Criminal networks are covert networks that operate in secrecy and stealth [8]. Criminals may minimize interactions to avoid attracting police attention and their interactions are hidden behind various illicit activities. Thus, data about criminals and their interactions and associations is inevitably incomplete, causing missing nodes and links in networks [10].
- *Incorrectness.* Incorrect data regarding criminals' identities, physical characteristics, and addresses may result either from unintentional data entry errors or from intentional deception by criminals. Many criminals lie about their identity information when caught and investigated.
- *Inconsistency.* Information about a criminal who has multiple police contacts may be entered into law enforcement databases multiple times. These records are not necessarily consistent. Multiple data records could make a single criminal appear to be different individuals. When seemingly different individuals are included in a network under study, misleading information may result.

Problems specific to criminal network analysis lie in data transformation, fuzzy boundaries, and network dynamics:

- *Data transformation.* Network analysis requires that data be presented in a specific format, in which network members are represented by nodes, and their associations or interactions are represented by links. However, information about criminal associations is usually not explicit in raw data. The task of extracting criminal associations from raw data and transforming them to the required format can be fairly labor-intensive and time-consuming.
- *Fuzzy boundaries.* Boundaries of criminal networks are likely to be ambiguous. It can be quite difficult for an analyst to decide whom to include and whom to exclude from a network under study [10].
- *Network dynamics.* Criminal networks are not static, but are subject to changes over time. New data and even new methods of data collection may be required to capture the dynamics of criminal networks [10].

Some techniques have been developed to address these problems. For example, to improve data correctness and consistency, many heuristics are employed in the FinCEN system at the U.S. Department of the Treasury to disambiguate and consolidate financial transactions into uniquely identified individuals in the system [5]. Other approaches like the concept space method [3] can transform crime incident data into a networked format [12].

Criminal Network Analysis and Visualization Tools

Klerks [7] categorized existing criminal network analysis approaches and tools into three generations.

First generation: Manual approach. Representative of the first generation is the Anacapa Chart [6]. With this approach, an analyst must first construct an association matrix by identifying criminal associations from raw data. A link chart for visualization purposes can then be drawn based on the association matrix. For example, to map the terrorist network containing the 19 hijackers in the September 11 attacks, Krebs [8] gathered data about the relationships among the hijackers from publicly released information reported in several major newspapers. He then manually constructed an association matrix to integrate these relations [8] and drew a network representation to analyze the structural properties of the network (Figure 1).

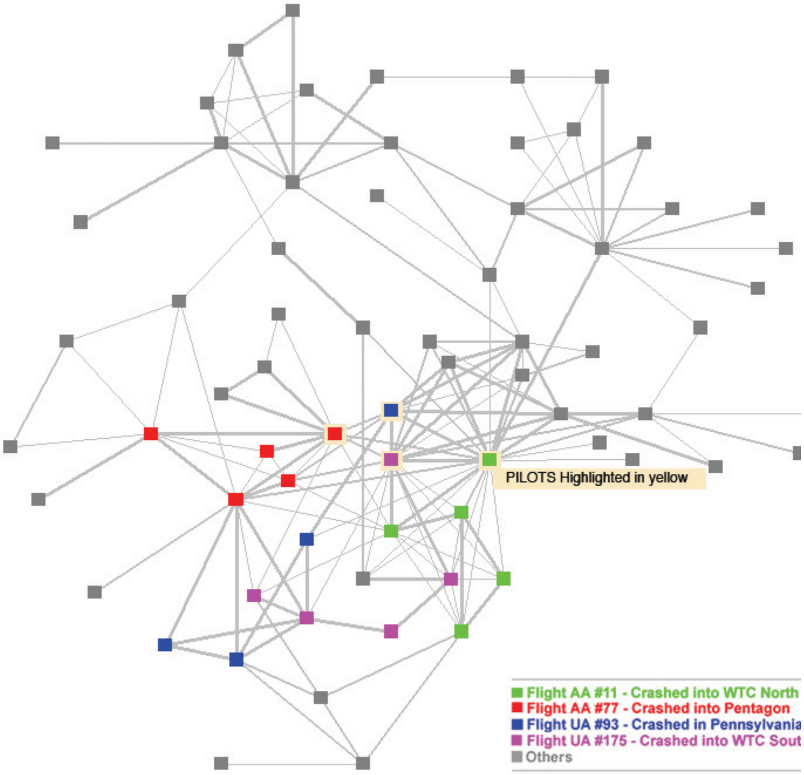

PILOTS Highlighted in yellow
🟩 Flight AA #11 - Crashed into WTC North
🟥 Flight AA #77 - Crashed into Pentagon
🟦 Flight UA #93 - Crashed in Pennsylvania
🟪 Flight UA #175 - Crashed into WTC Sout
■ Others

Figure 1. The terrorist network containing the 19 hijackers on Sept. 11, 2001.

Although such a manual approach for criminal network analysis is helpful in crime investigation, it becomes an extremely ineffective and inefficient method when data sets are very large.

Second generation: Graphic-based approach. These tools can automatically produce graphical representations of criminal networks. Most existing network analysis tools belong to this generation. Among them Analyst's Notebook [7], Netmap [5], and XANALYS Link Explorer (previously called Watson) [1], are the most popular. For example, Analyst's Notebook can automatically generate a link chart based on relational data from a spreadsheet or text file (Figure 2a).

Recently, two second-generation network analysis approaches have been developed by the COPLINK research and its co-developer, the Knowledge Computing Corporation. The first approach employs a hyperbolic tree metaphor to visualize crime relationships [3]. It is especially helpful for visualizing a large amount of relationship data because it simultaneously handles both focus and context (Figure 2b). The second approach uses a spring embedder algorithm [4] to adjust positions of nodes automatically to prevent a network display from being too cluttered. In such a view, icons represent different types of entities. A filtering function allows a user to select only those entity types of interest. (Figure 2c–e).

Although second-generation tools are capable of using various methods to visualize criminal networks, their sophistication level remains modest because they produce only graphical representations of criminal networks without much analytical functionality. They still rely on analysts to study the graphs with awareness to find structural properties of the network.

Third generation: SNA. This approach is expected to provide more advanced analytical functionality to assist crime investigation. Sophisticated structural analysis tools are needed to go from merely drawing networks to mining large volumes of data to discover useful knowledge about the structure and organization of criminal networks.

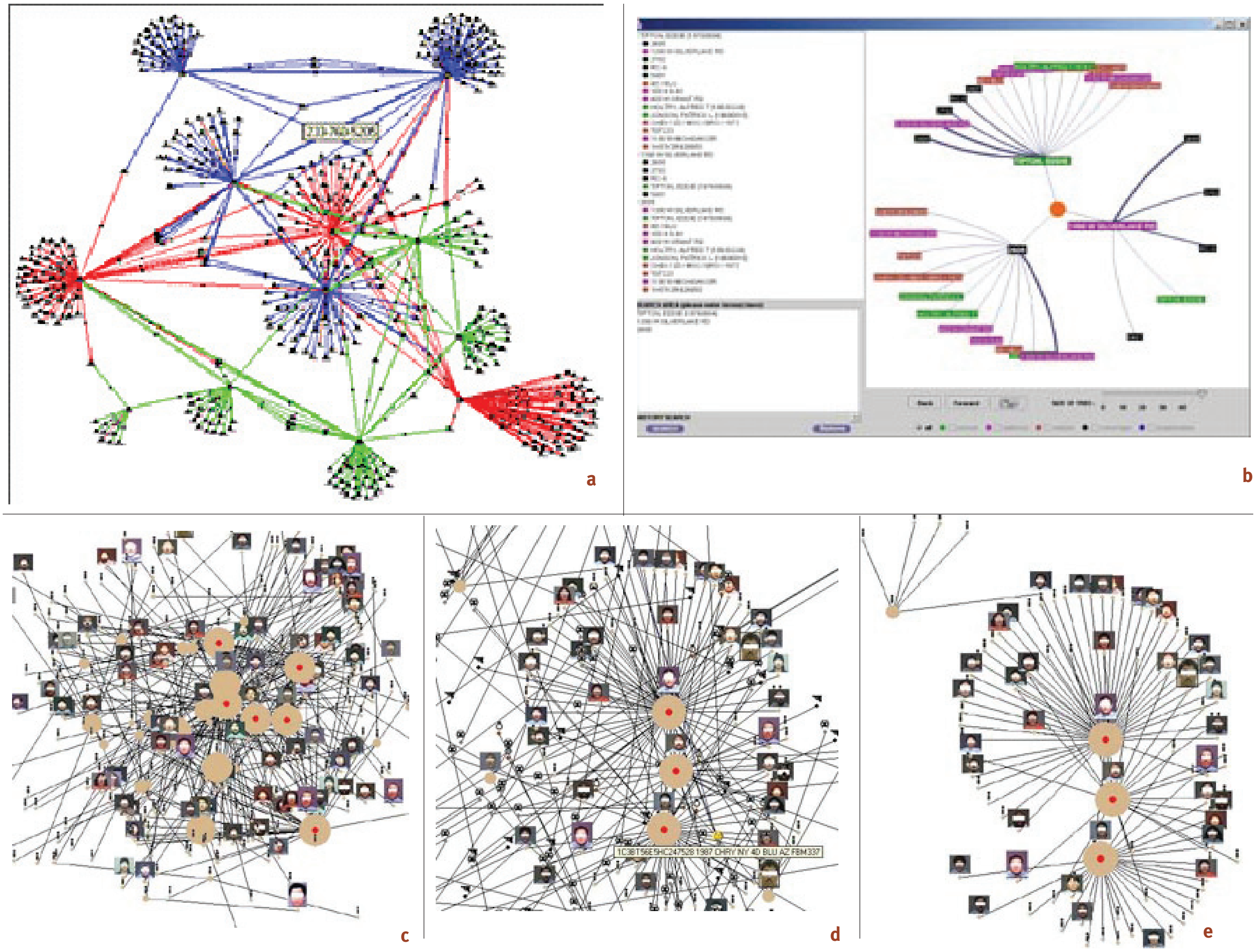


Figure 2. Second-generation criminal network analysis and visualization tools. (a) Analyst's Notebook (i2 Inc.; www.i2inc.com). The system can automatically arrange network nodes and allows one to drag nodes around for easier interpretation. (b) The hyperbolic tree view of relations among multiple criminal entities. (c) The initial layout of a criminal network produced by the network view (COPLINK Knowledge Computing Corp.; www.knowledgecc.com). (d) The network layout is adjusted automatically. The system moves the nodes having the largest number of links to the center of the display. A user can actually see movements of the nodes while their positions are adjusted and can fix the display at any time during the adjustment. (e) A user may choose only the entity type of interest (for example, person) and view textual explanations (for example, person name, address, and relations).

Data Mining Perspective

Intelligence and law enforcement agencies are often interested in finding structural properties of criminal networks [9]:

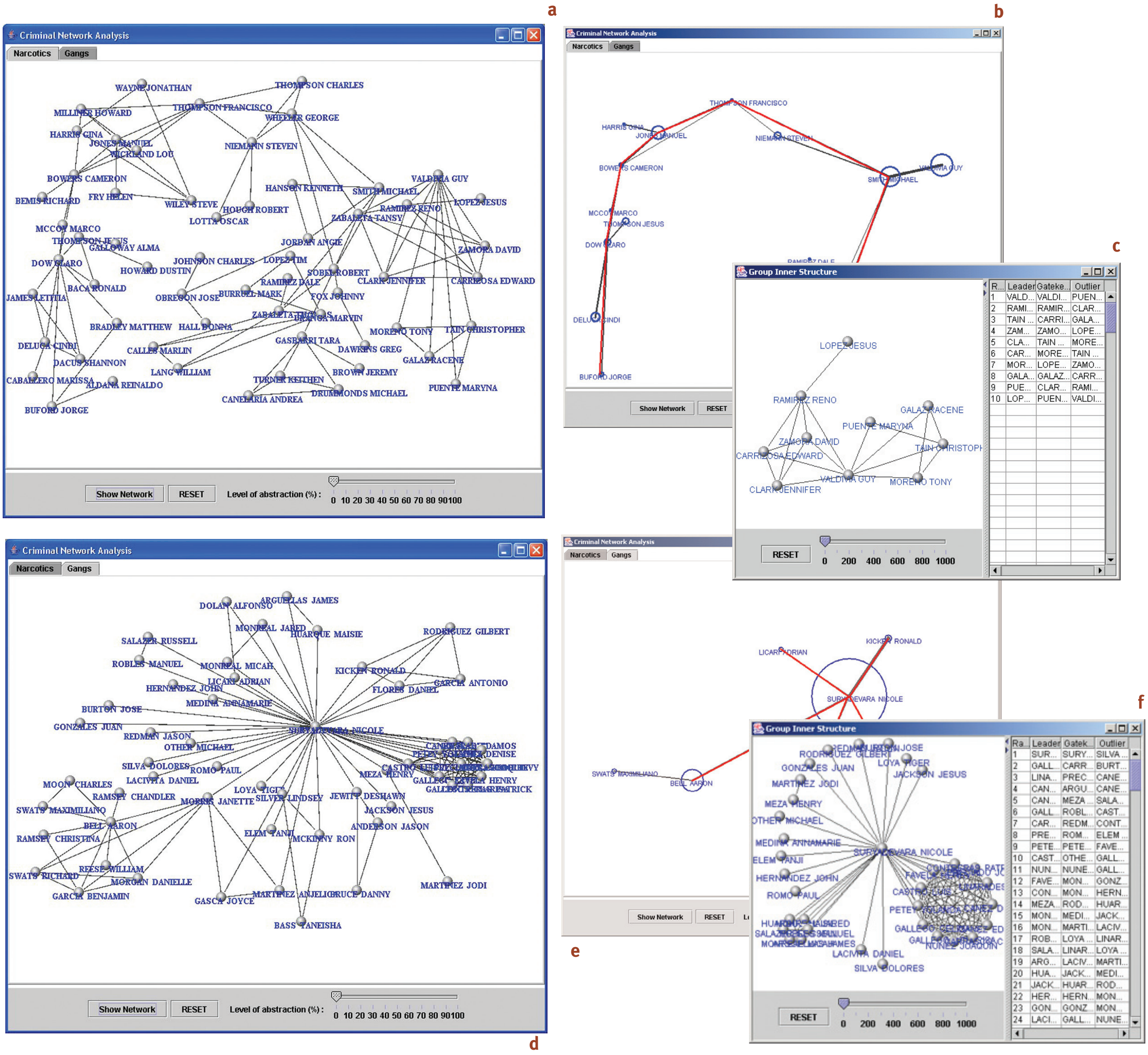
- What subgroups exist in the network?
- How do these subgroups interact with each other?
- What is the overall structure of the network?
- What are the roles (central/peripheral) network members' play?

A clear understanding of these structural properties in a criminal network may help analysts target critical network members for removal or surveillance, and locate network vulnerabilities where disruptive actions can be effective. Appropriate network analysis techniques, therefore, are needed to mine criminal networks and gain insight into these problems.

SNA. Because SNA techniques are designed to discover patterns of interaction between social actors in social networks [11], they are especially appropriate for studying criminal networks [9, 10].

Specifically, SNA is capable of detecting subgroups, discovering their patterns of interaction, identifying central individuals, and uncovering network organization and structure.

Subgroup detection. A criminal network can often be partitioned into subgroups consisting of individuals who closely interact with each other. Given a network, traditional data mining techniques such as cluster analysis may be employed to detect underlying groupings that are not otherwise apparent in



R...	Leader	Gateke...	Outlier
1	VALD...	VALDI...	PUEN...
2	RAMI...	RAMIR...	CLAR...
3	TAIN ...	CARRI...	GALA...
4	ZAM...	ZAMO...	LOPE...
5	CLA...	TAIN ...	MORE...
6	CAR...	MORE...	TAIN ...
7	MOR...	LOPE...	ZAMO...
8	GALA...	GALAZ...	CARR...
9	PUE...	CLAR...	RAMI...
10	LOP...	PUEN...	VALDI...

Ra...	Leader	Gatek...	Outlier
1	SUR...	SURY...	SILVA ...
2	GALL...	CARR...	BURT...
3	LINA...	PREC...	CANE...
4	CAN...	ARGU...	CANE...
5	CAN...	MEZA ...	SALA...
6	GALL...	ROBL...	CAST...
7	CAR...	REDM...	CONT...
8	PRE...	ROM...	ELEM ...
9	PETE...	PETE...	FAVE...
10	CAST...	OTHE...	GALL...
11	NUN...	NUNE...	GALL...
12	FAVE...	MON...	GONZ...
13	CON...	MON...	HERN...
14	MEZA...	ROD...	HUAR...
15	MON...	MEDI...	JACK...
16	MON...	MARTI...	LACIV...
17	ROB...	LOYA ...	LINAR...
18	SALA...	LINAR...	LOYA ...
19	ARG...	LACIV...	MARTI...
20	HUA...	JACK...	MEDI...
21	JACK...	HUAR...	ROD...
22	HER...	HERN...	MON...
23	GON...	GONZ...	MON...
24	LACI...	GALL...	NUNE...

Figure 3. The interaction patterns (and overall structures of networks) discovered from two criminal networks. (a) The network consists of 60 criminals dealing with narcotic drugs. It is difficult to manually detect subgroups and interaction patterns from this original network. (b) A chain structure becomes apparent using clustering and blockmodeling (see the red mark). Circles represent groups, which are labeled by their leaders' names, and straight lines represent between-group relationships. (c) The system can also show the inner structure of a selected group, identify its central members (leaders by degree, gatekeepers by betweenness, and outliers by closeness), and presents the centrality rankings of the members in a table in a separate window. (d) The network consisting of 57 gang members. (e) The star structure found in the gang network. (f) The details of a selected group in the gang network.

the data. Hierarchical clustering methods have been proposed to partition a network into subgroups [11]. Cliques whose members are fully or almost fully connected can also be detected based on clustering results.

Discovery of patterns of interaction. Patterns of interaction between subgroups can be discovered using an SNA approach called blockmodeling [11], which was originally designed to interpret and validate theories of social structures. When used in criminal network analysis, it can reveal patterns of between-group interactions and associations and can help reveal the overall structure of criminal networks under investigation. Given a partitioned network, blockmodel analysis determines the presence or absence of an association between a pair of subgroups based on a link density measure. In a network with undirected links, for example, the link density between two subgroups *i* and *j* can be calculated by d_{ij} = $n_i n_j$, where m_{ij} is the actual number of links between subgroups *i* and *j*; n_i and n_j represent the number of nodes within subgroups *i* and *j*, respectively. When the density of the links between the two subgroups is greater than a predefined threshold value, a between-group association is present, indicating the two subgroups interact with each other constantly and thus have a strong association. By this means, blockmodeling summarizes individual interaction details into interactions between groups so the overall structure of the network becomes more prominent.

Centrality deals with the roles of individuals in a network. Several centrality measures, such as degree, betweenness, and closeness can suggest the importance of a node in a network [11]. The degree of a particular node is its number of 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. An individual's having a high degree, for instance, may imply his leadership; whereas an individual with a high betweenness may be a gatekeeper in the network. Baker and Faulkner [2] employed these three measures, especially degree, to find the central individuals in a price-fixing conspiracy network in the electrical equipment industry. Krebs [8] found that in the network consisting of the 19 hijackers, Mohamed Atta scored the highest on degree and closeness, but not on betweenness.

Effective use of SNA techniques to mine criminal network data can have important implications for crime investigations. The knowledge gained may aid law enforcement agencies fighting crime proactively.

Implications. Effective use of SNA techniques to mine criminal network data can have important implications for crime investigations. For example, clustering with blockmodeling can help show the hidden structure of a criminal network. The knowledge gained may aid law enforcement agencies fighting crime proactively, for example, allocating an appropriate amount of police effort to prevent a crime taking place, or ensuring a police presence when the crime is carried out [9]. Sometimes, new structures discovered may even modify investigators' conventional views of certain crimes. For instance, Klerks [7] has found the stereotypical impression of hierarchical organizations within organized crime is being replaced by an image of more fluid and flattened networks. Traditional police strategies targeting leaders of a hierarchical criminal organization may have become less effective in fighting organized crimes today. The work by Krebs also demonstrates that the network consisting of the 19 hijackers in the September 11 attacks is fairly flat and dispersed [8]. The advantage of such a structure is an increase in the network's resilience and an emphasis on minimizing damage should some network members be captured or compromised.

SNA may also help address the challenges of data processing. Blockmodeling, for example, can easily detect "structural holes" [7] in which the link density is lower than a threshold density value. According to McAndrew [9], structural holes may indicate incomplete or missing data thereby drawing analysts' attention to further data collection and improvement.

The COPLINK Research Test Bed

Several data mining projects in the COPLINK research have begun to employ these SNA techniques for criminal network analysis. The goal has been to provide law enforcement and intelligence agencies with third-generation network analysis techniques that not only produce graphical representations of criminal networks but also provide structural analysis functionality to facilitate crime investigations. Prior to these data mining activities, several methods were employed to address the challenges of data processing. For inconsistency and incorrectness problems, we used the record linkage algorithm to relate multiple database records that actually refer to a single individual. For data transformation, we used the concept space approach [3] to extracting criminal associations from crime incident data.

The first stage of our network analysis development was intended to automatically identify the strongest association paths, or geodesics, between two or more network members using shortest-path algorithms. In practice, such a task often entails crime analysts to manually explore links and try to find association paths that might be useful for generating investigative leads. In the user study, our domain expert evaluated the paths identified automatically and those identified manually. He considered the former to be useful around 70% of time and the latter to be useful around only 30% of time.

Extending this attempt, a more sophisticated system for mining criminal network data has been developed. In addition to the visualization functionality, the system is intended to help detect subgroups in a network, discover interaction patterns between groups, and identify central members in a network.

Based on the crime incident data provided by the Tucson Police Department, several networks consisting of criminals involved in different types of crimes have been created and analyzed. Several domain experts validated the results of analysis (for example, subgroups, leaders, and gatekeepers). Moreover, interesting patterns of interactions between criminal groups and network structures were revealed in the networks. For example, a network of criminals dealing with narcotics and a network of gang members were found to have different structures, which became evident after cluster and blockmodel analysis (Figure 3). It appears the chain-structure network (for example, the narcotics network), may be disrupted by removing any part of the chain. Removing the central members from a star-structure network (for example, the gang network) might cause fatal damage to it.

To evaluate the system's performance we conducted a laboratory experiment involving 30 (student) subjects who performed 14 investigative tasks under two experimental conditions: structural analysis plus visualization (characteristics of third-generation tools), and visualization only (characteristics of second-generation tools). These 14 tasks were divided into two types: identifying interaction patterns between subgroups and identifying central members within a given subgroup. Our main performance metrics were effectiveness (defined as the total number of correct answers a subject generated for a given type of tasks) and efficiency (defined as the average time a subject spent to complete a given type of tasks). The results showed the average time spent on interaction pattern identification tasks under condition (a) (7.13 seconds) was significantly shorter than that under condition (b) (12.10 seconds; $t = 6.92$, $p < 0.001$). The difference in efficiency was also significant for central member identification tasks (6.24 seconds vs. 26.93 seconds; $t = 10.66$, $p < 0.001$). Such a gain in efficiency has important implications because time is of the essence for law enforcement and intelligence agencies seeking to prevent or respond to terrorist attacks or other serious crimes. No significant improvement in effectiveness was present ($t = 1.80$, $p > 0.05$), probably due to the small sizes and relatively simple structures of the testing networks [12].

Conclusion

It is believed that reliable data and sophisticated analytical techniques are critical for law enforcement and intelligence agencies to understand and possibly disrupt terrorist or criminal networks. Using automated SNA and visualization techniques to reveal various structures and interactions within a network is a promising step forward. The continued advancement of criminal network analysis techniques will enable us finally to win the new netwar. C

References

1. Anderson, T., Arbetter, L., Benawides, A., and Longmore-Etheridge, A. Security works. *Security Management 38,* 17, (1994), 17–20.
2. Baker, W.E. and Faulkner, R.R. The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry. *Amer. Sociological Rev. 58,* 2 (1993), 837–860.
3. Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W., and Schroeder, J. COPLINK: Managing law enforcement data and knowledge. *Commun. ACM 46,* 1 (Jan. 2003), 28–34.
4. Eades, P. A heuristic for graph drawing. *Congressus Numerantium 42* (1984), 149–160.
5. Goldberg, H.G., and Senator, T.E. Restructuring databases for knowledge discovery by consolidation and link formation. In *Proceedings of 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis.* AAAI Press (1998).
6. Harper, W.R., and Harris, D.H. The application of link analysis to police intelligence. *Human Factors 17*, 2 (1975), 157–164.
7. Klerks, P. The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. *Connections 24,* 3 (2001), 53–65.
8. Krebs, V. E. Mapping networks of terrorist cells. *Connections 24*, 3 (2001), 43–52.
9. McAndrew, D. The structural analysis of criminal networks. *The Social Psychology of Crime: Groups, Teams, and Networks, Offender Profiling Series, III.* D. Canter and L. Alison (Eds.). Aldershot, Dartmouth (1999).
10. Sparrow, M.K. The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks 13* (1991), 251–274.
11. Wasserman, S., and Faust, K. *Social Network Analysis: Methods and Applications.* Cambridge University Press, Cambridge, MA, 1994.
12. Xu, J., and Chen, H. CrimeNet explorer: A framework for criminal network knowledge discovery. To appear in *ACM Trans. on Info. Systems.*

This work has primarily been funded by the National Science Foundation, Digital Government Program COPLINK Center: Information and Knowledge Management for Law Enforcement, #9983304, July, 2000–June, 2003.

Jennifer Xu (jxu@eller.arizona.edu) is a doctoral candidate in management information systems at the University of Arizona, Tucson, AZ.

Hsinchun Chen (hchen@eller.arizona.edu) is McClelland Professor of Management Information Systems and head of Artificial Intelligence Lab at the MIS Department at the University of Arizona, Tucson, AZ.