

Cross-Jurisdictional Criminal Activity Networks to Support Border and Transportation Security

Byron Marshall, Siddharth Kaza, Jennifer Xu, Homa Atabakhsh, Tim Petersen, Chuck Violette, and Hsinchun Chen

Abstract—Border and transportation security is a critical part of the Department of Homeland Security’s (DHS) national strategy. DHS strategy calls for the creation of “smart borders” where information from local, state, federal, and international sources can be combined to support risk-based management tools for border-management agencies. This paper proposes a framework for effectively integrating such data to create cross-jurisdictional Criminal Activity Networks (CAN)s. Using the approach outlined in the framework, we created a CAN system as part of the DHS-funded BorderSafe project. This paper describes the system, reports on feedback received from investigating officers, and highlights key issues and challenges.

I. INTRODUCTION

IN the aftermath of the Sept. 11th attacks, a National Strategy for Homeland Security was developed [1] to help guide government and private-sector security efforts. This strategy identifies “border and transportation security” and “protecting critical infrastructures and key assets” as two of the six critical mission areas. The report calls for the creation of “smart borders” which provide “greater security through better intelligence, coordinated national efforts, and unprecedented international cooperation against the threats posed by terrorists, the implements of terrorism, international organized crime, illegal drugs, illegal migrants, cyber crime, and the destruction of natural resources [1] p22.” The plan calls upon the Department of Homeland Security (DHS) to provide increased information on inbound goods and passengers to support risk-based management tools for border-management agencies.

While monitoring millions of border crossings each year,

DHS balances operational efficiency and security concerns. Thorough vehicle checks are important for public safety but if the lines at the border become too long the flow of people and vehicles is impaired. As vehicles enter the country, DHS records each license plate with a crossing date and time. Agents also search vehicles for drugs and other contraband. The value of the collected data could be enhanced by combining it with the millions of relationships recorded between people, places, vehicles, and incidents in local law enforcement records management systems (RMS), but usefully integrating the data is a difficult task. In spite of an obvious connection between criminal and border crossing activities, such data is only combined in special cases when investigators call each other and ask for help. While investigators are generally cooperative in these matters, this kind of sharing is time-consuming and therefore infrequent.

Data sharing has important implications both for homeland security and local law enforcement. Suspicious activity and other reports from locations near critical infrastructure sites can have national security implications. Local law enforcement officials may have data related to terrorists without knowing the individuals are terrorists. Border agencies are interested in certain individuals but have no efficient way to check with local authorities. This overlap of data is increasingly important because drug traffickers are thought to be “branching out” into the smuggling of illegal aliens across the border. Occasional reports of criminal networks in a foreign country attempting to influence U.S. criminal justice activities are also a concern. U.S. Attorney Paul Charlton recently noted two such cases. In one case eight witnesses were assassinated by a Mexican prison gang and in another an assistant state attorney was targeted by a gunman [2].

This work develops a methodology for identifying important investigative leads by analyzing known relationships between people, vehicles, criminal incidents, and border-crossing activity. We explore the use of cross-jurisdictional criminal activity network (CAN) evaluation in support of border transportation and safety investigations. In CAN analysis, details such as locations, physical descriptions, and known associations are organized into networks designed to help identify

Manuscript received April 1, 2004. This work was supported in part by the NSF, Knowledge Discovery and Dissemination (KDD) # 9983304, June 2003-March 2004, NSF, ITR: “COPLINK Center for Intelligence and Security Informatics Research - “A Crime Data Mining Approach to Developing Border Safe Research.” Sept. 1, 2003 - Aug. 31, 2004, Department of Homeland Security (DHS) / Corporation for National Research Initiatives (CNRI): “Border Safe,” Sept. 2003 - Nov. 2004

Byron Marshall, Siddharth Kaza, Jennifer Xu, Homa Atabakhsh, and Hsinchun Chen are with the AI Lab in the University of Arizona’s Dept. of MIS. phone: 520-621-3927; fax: 520-621-2433; e-mail: byronm@eller.arizona.edu. Tim Petersen and Chuck Violette are with the Tucson Police Department.

investigative “leads”. Our system was developed to explore two questions:

- Does cross-jurisdictional information increase the value of CANs in identifying investigative leads?
- How can existing sources be exploited and what are the major obstacles to be faced in creating cross-jurisdictional criminal activity networks?

In the next section we review background information. Section III presents a framework for constructing cross-jurisdictional CANs and Section IV describes our research testbed and system implementation. Section V discusses the work and includes our conclusions and future directions.

II. LITERATURE REVIEW

Combining data from independently-developed sources is a challenging task. Information integration approaches such as federation, warehousing, and mediation aim to address different needs and difficulties [3]. Commonly acknowledged problems [4] include (1) Name Differences: same name, different entity, (2) Mismatched Domains: problems with units of measure or reference point, (3) Missing Data: incomplete sources or different data available from different sources, and (4) Object Identification: no global ID values and no inter-database ID tables.

The task of integrating two databases can be generally divided into two parts [5]. First schema-level heterogeneity is resolved by aligning semantically corresponding columns between the two sources. Secondly, entity matches are identified to connect objects in one database to records describing the same objects in the other database. Entity level matching is generally performed after schema-level matching is complete. The nature of the overlap between different datasets strongly affects the entity matching process. Existing matching processes can be categorized as using (1) key equivalence, (2) user specified equivalence, (3) probabilistic key equivalence, (4) probabilistic attribute equivalence, or (5) heuristic rules [5]. Variations of these approaches can be seen in existing law enforcement data sharing initiatives.

Law enforcement personnel are confident that cross-jurisdictional data sharing information is important but it is difficult to combine information from more than one jurisdiction for several reasons. Most law enforcement RMS systems are not interoperable. Because of unique needs, policy implications, cultural momentum, and existing contractual arrangements it would be very expensive and organizationally difficult to get multiple jurisdictions to use compatible systems. Even if systems are quite similar in structure and function, identity records are difficult to match across jurisdictions. Finally, because vast quantities of data exist in each local system, combining several systems could reduce performance and effectiveness.

In the law enforcement domain schema-level heterogeneity is generally addressed using one of the three approaches described here.

1. Agencies allow remote access to existing systems. This first approach emphasizes connectivity. For example, the ARJIS system [6] is an extensive network accessing information from a large number of San Diego area criminal justice agencies. For these systems, data integration involves mediators that consolidate and translate queries to gather results from multiple sources.

2. Source data is mapped to other structures to support specific tools. Implementations based on this approach emphasize specific desired functionality. One example is the COPLINK system which implements an investigation-oriented database structure to support queries over law enforcement data. COPLINK was initially developed at the University of Arizona’s AI Lab [7, 8] and a commercial version is now developed and distributed by Knowledge Computing Corporation (KCC) [9]. KCC creates migration routines to extract data from a client agency RMS and organize them into a flexible database structure called a COPLINK “Node”. When several agencies deploy COPLINK it can be configured to support cross-jurisdictional data searching.

3. Standardized data structures are evolving to formalize the semantics of available data, as in the GJXDM project [10]. This approach is intended to decrease the cost and time required to implement data sharing between agencies. This is an attractive approach but an extensive set of standardized objects has not yet been widely accepted.

Even when the technical issues have been addressed, policy and privacy issues remain. Installations that host criminal activity information need to take special care to prevent unintended release of data. Investigators do not want targets of investigation to know they are being watched and smuggling organizations are known to respond quickly to changes in border monitoring activity. Thus, sharing of data between different agencies requires customized data sharing agreements [11].

Criminal Activity Networks are frequently depicted in manually produced link charts developed by law enforcement personnel to support crime analysis. Link charts include important individuals and relationships discovered in the course of criminal investigation. Several computerized tools have been developed to support link-chart-like representations of criminal activity information including NetMap, Analyst’s Notebook, and COPLINK’s visualizer. NetMap processes large collections of associations by decomposing the data into nodes and links and generating charts that use a line’s thickness or color to annotate associations between nodes [12]. Analyst’s Notebook from i2 supports analysis and visualization of networks of criminal activity [13]. i2 has created tools to store manually input data and map to existing external

databases. Another example is KCC’s COPLINK visualizer component which displays relationships and supports user drill-down to underlying details [14]. Various information features are depicted by this visualization tool. Relational closeness is reflected in close proximity and levels of activity are reflected in icon size. These tools provide various levels of interaction and pattern identification.

III. CROSS-JURISDICTIONAL INTEGRATION FRAMEWORK

Our review of existing systems and domain-specific considerations helped us develop the following framework for creating cross-jurisdictional information CANs. The key to the framework is identification of 3 classes of data: (1) **base data** with overlapping data from multiple jurisdictions with multiple object and relation types, (2) high volume but relatively simple **supplementary data** to enhance CAN information content, and (3) case specific or ad-hoc **query-specific** data expressing important relationships or features. Given these classes of data, integration should proceed in three steps: schema-level transformation of base data, entity-matching to align objects across data sets, and normalization and matching of supplementary data.

Base data should be semantically aligned and mapped to support CAN generation. When there is a lot of overlap between datasets, there is a lot of value to be gained. This is a classic data integration task requiring reconciliation of legacy data into a common schema and instance-level entity matching. Police RMS records are the prime example of this kind of data because multiple jurisdictions keep similar types of data about an overlapping set of objects. Standardized data dictionaries may eventually encourage development of interoperable systems, but for now data sharing initiatives generally begin by mapping to a global schema and then move on to entity matching. Base data integration should be a repeatable transformation process so that the combined datasets can be refreshed frequently.

Entity matching in this domain will tend to rely on heuristics. Primary objects will include people, locations, and vehicles. More research into appropriate identity matching algorithms for cross-jurisdictional datasets is needed. Previous and current work in the AI Lab aims to address this issue [15]. Input from domain experts suggests an initial match for people using first name, last name, and date of birth. These heuristics are not perfect; a few incorrect matches may result and certainly many correct matches will be missed. Other alternatives such as FBI and state numbers may be useful but are not consistently available. Locations can be matched based on geo-codes and vehicles can be matched by license plate and/or vehicle identification number (VIN).

License plate data has some interesting and useful characteristics. Plate numbers can be recorded in an unobtrusive fashion and, while criminals frequently avoid identification by lying about their names in routine

interactions with law enforcement officials, license plate numbers are directly observed. In addition, vehicles used by criminals are often registered in someone else’s name. Even if a criminal uses an alias in incidents involving a particular vehicle, the resulting person-vehicle data implicitly links the incidents. License plate numbers also are occasionally transferred to different cars: illegally when a car or plate is stolen or legally when it is sold. For many applications these characteristics make plate numbers more useful than vehicle identification numbers.

In addition to the base data, investigators use many additional supplementary or query-specific information resources to identify criminals’ activities and associations. This additional data may not be readily available for a variety of reasons.

- **Specialization:** Frequently, useful data is not directly accounted for in the global schema. For example, police RMS systems do not usually store border-crossing events.
- **Availability:** Frequently, information like jail visitation histories and motor vehicle registration records are important and could be, but haven’t been, included in an agency’s data system.
- **Sensitivity:** Investigators do not want many bits of information included in widely used sources. In some cases it is feared that information would be leaked to the criminals involved. In some cases data has been subpoenaed and can be used only in a single investigation.
- **Contextual usefulness:** Background information and rumors identify some relationships between individual criminals, for example, “Bob and Joe are brothers” or “Fred and Jim were friends in high school”. This kind of information is not collected in large quantities, applies only to specific cases, and should not be included in an RMS because of privacy and security concerns.

Our framework allows for the inclusion of this kind of data by treating it as supplementary data or as query-specific data. A data source is appropriate for supplementary integration when (1) it is available in quantity and can be appropriately organized, (2) its sensitivity level allows for it to be shared across multiple investigations, and (3) it is contextually appropriate outside of a single investigation. Data can be used as supplementary data if it can be reduced to one or more lists of features or events directly associated with identifiable objects in the base data set. For example, mug shots of people, border crossing records, or jail visitations can all be recorded associated with particular individuals already contained in a base data set of criminal incidents. Query-specific data can be used to guide CAN building. For example if phone records indicate a suspect called 19 different people, a CAN network could query for relationships involving any of the 20 people to arrive at a more context-specific result without storing subpoenaed data in the general investigation data set. Both

supplementary and query-specific data has to be normalized and matched to the objects and entities from the base data.

IV. RESEARCH TESTBED: BORDERSAFE

The BorderSafe project (funded by the Department of Homeland Security) is a collaborative research effort involving the University of Arizona's AI Lab, law enforcement agencies including the Tucson Police Department (TPD), Phoenix Police Department (PPD), Pima County Sheriff's Department (PCSD), and Tucson Customs and Border Protection (CBP) as well as San Diego ARJIS (Automated Regional Justice Information Systems) and the San Diego Super Computer Center (SDSC). The BorderSafe project includes several cross-jurisdictional data sharing initiatives.

We integrated TPD and PCSD datasets with each other and with a dataset made available for this research by CBP. Table I describes the TPD and PCSD datasets. The TPD and PCSD jurisdictions cover adjacent areas in and around the city of Tucson, AZ, with a combined population approaching one million people. In many ways the two jurisdictions represent a shared community of citizens. They also share intertwined communities of criminals. The CBP dataset was handled as a supplementary source that identifies border crossing vehicles. Video equipment automatically extracts license plate numbers of cars as they cross into or out of the country. Camera errors are manually corrected as incoming vehicles pass through the port of entry. CBP shared a selected data set with the BorderSafe project (see Table II).

TABLE I KEY STATISTICS FROM THE TPD AND PCSD DATASETS

	TPD	PCSD
Recorded Incidents	2.84 million	2.18 million
Persons	1.35 million	1.31 million
Vehicles	623,656	520,539

TABLE II CBP BORDER CROSSING DATA

1,125,155	Records: plate, state, date, & time
226,207	Distinct vehicles
209	Days of information out of 18 months
130,195	Plates issued in AZ
5,546	Plates issued in CA
90,466	Plates issued in Mexico

A. Integrating the Data

Integration of the data sets proceeded in three steps:

1. TBP/PCSD records were mapped to a common schema.
2. Cross-jurisdictional identities were matched.
3. CBP data was imported as a supplementary source.

Because the TPD and PCSD data come from cooperative agencies with closely related activities, it was appropriate to invest in a significant integration effort. Automated transformation procedures now translate PCSD and TPD

RMS records into COPLINK format. People were matched on FirstName, LastName, and DateOfBirth (DOB). Each vehicle plate found in the police records was matched to the CBP crossing data to establish a border crossing history and help identify potentially interesting vehicles. We expect that future versions of our system will allow inclusion of other supplemental and query-specific data such as family associations, phone records, and jail visitations. These sources could be added as part of an interactive link-chart drawing process.

We began analyzing our data by evaluating the overlap between datasets. We tallied the number of out-of-state vehicles in the TPD records, finding 7% of the vehicles involved in gang-related, violent, and narcotics crimes are registered outside of Arizona. Table III shows the number of plates found in both the CBP crossing data and TPD/PCSD data sets, confirming that a lot of cross-border activity can be identified for vehicles connected to crime incidents.

TABLE III BORDER CROSSING PLATES BY DATASET

	TPD	PCSD	Combined
Border Crossing Plates found in Police Records	8,300	6,619	13,111
Crossing Associated with Those Vehicles	34,632	31,075	59,275

We also looked at person overlap. More than 483,000 people appear in both the TPD and PCSD datasets (36% of all records). It may well be that more overlap is masked by clerical errors, missing data, or intentional deception.

B. CAN Evaluation

Next we measured the impact of cross-jurisdictional information on activity networks traced in the criminal activity records. We randomly chose 50 people from a combined list of wanted suspects and known drug traffickers. We selected only people appearing in both TPD and PCSD records (the large majority did appear in both data sets). We used the associations or links that occur when individuals or vehicles are listed together in an incident report to trace CANs. Linkages such as shared connections to locations (like crackhouse addresses) could be useful but have not yet been implemented in our system.

For each person we followed all known person-to-person associations and compiled a list of people. Links for each person in the new list were also followed. We then followed person-to-vehicle links to identify plate numbers. The result was a network of all people within two "hops" of the focus individual and all associated vehicles known to have cross-border activity. We created three networks for each person: one with links from the TPD dataset, one with links found in the PCSD dataset, and one using the links in both datasets. Table IV reports the average number of associated people, associated vehicles, and associational links found for the 50 selected individuals. It is not

surprising that combining the data sets allowed us to connect more people and border crossing vehicles for this list of known criminals.

TABLE IV
THE IMPACT OF CROSS-JURISDICTIONAL DATA: AVERAGE NUMBER OF PEOPLE, VEHICLES, AND LINKS FOR 50 SELECTED INDIVIDUALS

	People	Vehicles	Associations
TPD associations only	193.3	1.44	659.92
PCSD associations only	120.36	1.08	1,016.14
All associations	389.92	6.92	2,487.14

Next we created a set of CAN visualizations for review by law enforcement personnel. We limited our networks to 50 nodes at most, because more nodes overwhelmed the viewer. Cognitive overload literature and experimentation should help establish appropriate initial network sizes for display. While the size of a network may “converge” quickly if little information is available, networks frequently become unmanageable in just a few iterations. A variety of visual cues were used in this preliminary implementation. We differentiated entity types by shape, key attributes by node color, degree of activity as node size, connection source by link color, and some details in link text or roll-over tool tips. Figure 1 shows a network connecting narcotics traffickers and border crossing plates.

- Associations found in the TPD data are blue, PCSD links are green, and associations noted in both sets are red.
 - Node size indicates the extent of criminal activity. All incidents involving a person are counted. Violent, narcotics-related, and gang-related activities are counted twice. The activity scores are normalized to identify the relative activity levels of the individuals in the network. Future work will explore various methods of determining appropriate node size.
 - Nodes in the display are initially arranged by a spring-embedder algorithm which “pushes” and “pulls” nodes using the links in the network. This algorithm needs further development but generally tends to place closely related people near each other in the display area.
- Choice of these features was guided by several intuitive notions gleaned from conversations with investigators: high levels of criminal activity and frequent border crossings signal useful investigative leads, crime types and person roles are important for association evaluation, and longer associative paths are less interesting.

We maintained close contact with law enforcement personnel to verify project assumptions, identify useful functions, and understand how networks can be used or evaluated. We implemented a system to enhance visualizations of TPD data with the CBP border crossing information and created networks of people and/or vehicles which all included at least one vehicle with recorded border crossings.

Police crime analysts and CBP personnel were shown networks like those depicted in Figures 1, 2 and 3. The

ensuing conversations generated a number of comments and suggestions which are summarized here.

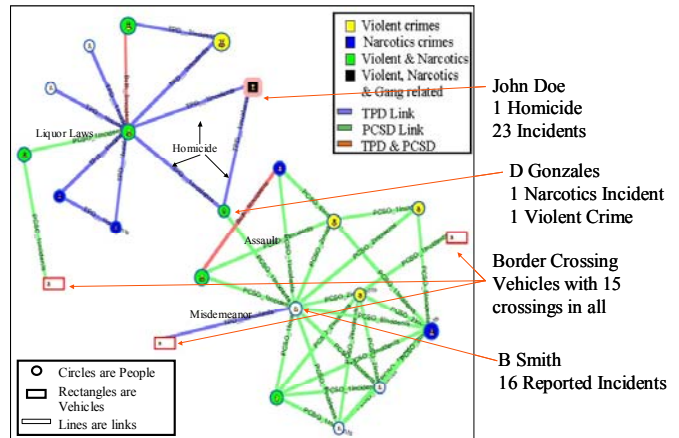


Fig. 1. Visualization of a Criminal Activity Network

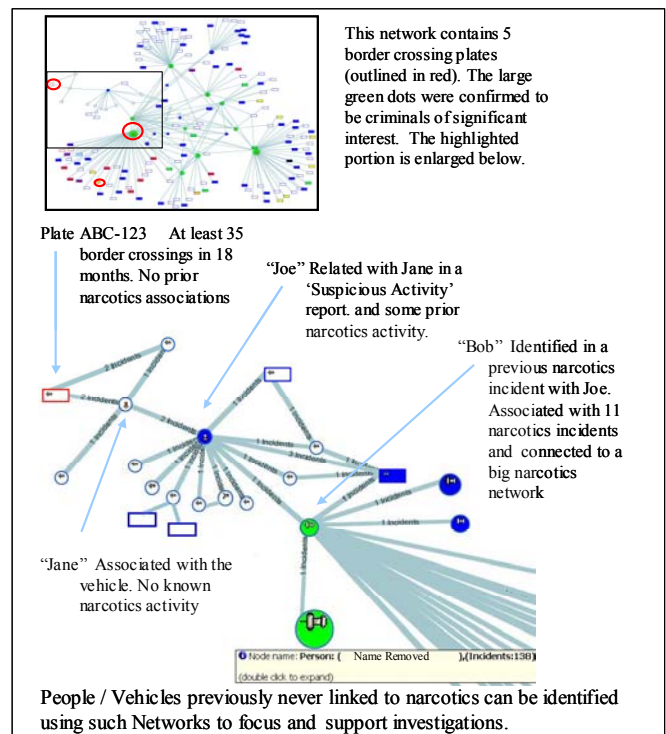


Fig. 2. A complex network connecting border crossing plates and known drug trafficker

These networks can be easily analyzed to reveal links between relatively obscure subjects who are routinely crossing the border and known participants in Tucson’s drug trade. That information could subsequently be used to focus and direct law enforcement resources and investigations. Automatically generated activity networks for wanted individuals would save a lot of time. A network display tool would be used frequently and in some large cases (3-4 times per year) such a tool could save 100 person hours on a single project.

Indications of cross-border activity would be very useful in focusing certain investigations. Correlating stolen

vehicle reports with border crossings and targeted individuals could help in many investigations, but finding correlations manually was very time consuming.

No one was surprised that activity networks associated with known traffickers and “most-wanted” individuals were substantially expanded by inclusion of links from both TPD and PCSD.

V. DISCUSSION

In addition to the comments summarized in the previous section, several promising border and transportation security scenarios were identified.

1. Police agencies can check vehicles associated with criminals against border-crossing records. In one on-going investigation, a suspect is reported to frequently cross the border. The dates of some of these visits are known. Investigators want to know which of the dozens of vehicles indirectly associated to the suspect have crossed the border in the given time frame.

2. CBP can adjust port operations based on vehicles' associations to known drug traffickers. For example, when a vehicle is searched and found to have drugs, vehicles crossing before and after are scrutinized to identify patterns of activity. Knowing that particular vehicles had indirect associations to drug criminals might help them identify the patterns of activity associated with drug smuggling.

3. Police agencies could use the plate numbers of vehicles found carrying drugs to focus on-going investigations.

4. Incidents near critical infrastructure sites can be organized to alert agents to patterns of activity. For example, a location-centered CAN would quickly highlight a vehicle appearing in suspicious activity reports near two or more critical infrastructure sites.

In each of these real life scenarios, border or law enforcement activities would be enhanced by the use of network-style, cross-jurisdictional association analysis. Increasing the information flow to the border agencies should allow them to make better operational decisions and the flow of information back to local agencies would be helpful in pursuing related investigations.

One outcome of the work so far is creation of an infrastructure to expand research related to CANs. On-going experiments aim to identify the features that make one CAN visualization more useful than another. Various investigational scenarios are also being explored, both by providing information to on-going investigations and in experimental work. This kind of evaluation is important to the development of selection algorithms that will automatically choose a high-value initial set of associations in response to a query.

ACKNOWLEDGMENT

The authors thank members of the University of Arizona Artificial Intelligence Lab, including Ben Smith and

Chunju Tseng (Lu) who spearheaded development of the visualization component, Tucson Police Department, Pima County Sheriff's Department, and Tucson Customs and Border Protection.

REFERENCES

- [1] "National Strategy for Homeland Security. U.S. Office of Homeland Security, July 2002; www.whitehouse.gov/homeland/book/."
- [2] M. Marizco, "U.S. official: Drug violence crossing border," in *Arizona Daily Star*. Tucson, 2003.
- [3] H. Garcia-Molina, J. D. Ullman, and J. Widom, *Database systems, the complete book*. Upper Saddle River, NJ: Prentice-Hall, 2002.
- [4] I.-M. A. Chen and D. Rotem, "Integrating information from multiple independently developed data sources," presented at Seventh international conference on information and knowledge management, Bethesda, Maryland, 1998.
- [5] E.-P. Lim, J. Srivastava, S. Prabhakar, and J. Richardson, "Entity identification in database integration," *Information Sciences*, vol. 89, pp. 1-38, 1996.
- [6] Automated Regional Justice Information System accessed 3/24/2004 <http://www.arjis.org/>.
- [7] H. Chen, J. Schroeder, R. V. Hauck, L. Ridgeway, H. Atabakhsh, H. Gupta, C. Boarman, K. Rasmussen, and A. W. Clements, "COPLINK connect: information and knowledge management for law enforcement," *Decision Support Systems*, vol. 34, pp. 271-285, 2002.
- [8] H. Chen, D. Zeng, H. Atabakhsh, W. Wyzga, and J. Schroeder, "COPLINK managing law enforcement data and knowledge," *Communications of the ACM*, vol. 46, pp. 28-34, 2003.
- [9] "Software Joins Cops on the Beat," COPLINK program links databases, speeds police investigations in the state of Alaska., in *Anchorage Daily News*.
- [10] "Office of Justice Programs, global justice XML data model," US Department of Justice 2004.
- [11] H. Atabakhsh, C. Larson, T. Petersen, C. Violette, and H. Chen, "Information sharing and collaboration policies within government agencies (forthcoming)," presented at 2nd Symposium on Intelligence and Security Informatics, June 10-11 2004, Tucson, AZ, 2004.
- [12] E. Chabrow, "Tracking The Terrorists: Investigative skills and technology are being used to hunt terrorism's supporters," in *Information Week*, 2002.
- [13] I2 Investigative Analysis Software accessed 3/24/2004 http://www.i2inc.com/Products/Analysts_Notebook/.
- [14] COPLINK from Knowledge Computing Corp. accessed 3/24/2004 <http://www.coplink.net/vis1.htm>.
- [15] G. Wang, H. Chen, and H. Atabakhsh, "Automatically detecting deceptive criminal identities," *Communications of the ACM*, vol. 47, pp. 70-76, 2004.