Topological Analysis of Criminal Activity Networks: Enhancing Transportation Security

Siddharth Kaza, Jennifer Xu, Byron Marshall, *Member, IEEE*, and Hsinchun Chen, *Fellow, IEEE*

***Abstract*—The security of border and transportation systems is a critical component of the national strategy for homeland security. The security concerns at the border are not independent of law enforcement in border-area jurisdictions because the information known by local law enforcement agencies may provide valuable leads that are useful for securing the border and transportation infrastructure. The combined analysis of law enforcement information and data generated by vehicle license plate readers at international borders can be used to identify suspicious vehicles and people at ports of entry. This not only generates better quality leads for border protection agents but may also serve to reduce wait times for commerce, vehicles, and people as they cross the border. This paper explores the use of criminal activity networks (CANs) to analyze information from law enforcement and other sources to provide value for transportation and border security. We analyze the topological characteristics of CAN of individuals and vehicles in a multiple jurisdiction scenario. The advantages of exploring the relationships of individuals and vehicles are shown. We find that large narcotic networks are small world with short average path lengths ranging from 4.5 to 8.5 and have scale-free degree distributions with power law exponents of 0.85–1.3. In addition, we find that utilizing information from multiple jurisdictions provides higher quality leads by reducing the average shortest-path lengths. The inclusion of vehicular relationships and border-crossing information generates more investigative leads that can aid in securing the border and transportation infrastructure.**

***Index Terms*—Border and transportation security, homeland security, social network analysis.**

I. Introduction

A NATIONAL strategy for homeland security [1] was developed after the September 11th terrorist attacks and presented in a report published by the Office of Homeland Security in July 2002. The report emphasizes "Border and Transportation Security" and "Protecting Critical Infrastructures and Key Assets" as two of the six critical mission areas. Transportation systems (both domestic and transnational) have been identified as key infrastructures that need to be protected. In addition, the report calls for the creation of "smart borders" that provide "greater security through better intelligence, coordinated national efforts, and unprecedented international cooperation [1]."

Information needed for securing transportation systems exists in multiple agencies that frequently do not share information horizontally (across the same level of government) or vertically (across local, state, and federal government). If homeland-security-related information were efficiently shared, then border and infrastructure protection would be benefited. Information sharing can also help improve traffic flow at the border while balancing security concerns. The identification of suspect vehicles at the border could be enhanced by combining it with the millions of relationships recorded between people, places, and vehicles in law enforcement records of border-area jurisdictions. Similarly, local law enforcement information would be of benefit to the Transport Security Administration (TSA) in protecting transportation infrastructures. Criminal activity networks (CAN) can be used to analyze information from multiple sources like law enforcement and transportation systems like license plate readers.

A CAN is a network of interconnected people (known criminals), vehicles, and locations based on law enforcement records. The networks can be augmented with data from sources like transportation systems and motor vehicle division data. These networks allow us to analyze and visualize information that is helpful for identifying suspicious vehicles and people at the border or around critical infrastructures. CANs may contain information from multiple sources and be used to identify relationships between people and vehicles that are unknown to a single jurisdiction. As a result, cross-jurisdictional information sharing and triangulation can help generate better investigative leads and strengthen legal cases against criminals.

CANs can be large and complex (particularly in a cross-jurisdictional environment) and can be better analyzed if we study their topological properties. Topological properties describe the network as a whole and help us better understand its governing mechanisms. The topological properties can also be used to quantify the advantages of data sharing to law enforcement and transportation security. In addition, understanding the properties of CANs can help design better analysis tools to assist in identifying potentially dangerous vehicles and people. In this paper, we study the topological properties

Manuscript received February 6, 2007; revised March 26, 2008 and October 3, 2008. First published February 2, 2009; current version published February 27, 2009. This work was supported in part by the National Science Foundation (NSF) Knowledge Discovery and Dissemination (KDD) program under Grant 9983304, by the NSF Information Technology Research (ITR) program "COPLINK Center for Intelligence and Security Informatics Research—A Crime Data Mining Approach to Developing Border Safe Research" under Grant 0326348, and by the Department of Homeland Security (DHS) and Corporation for National Research Initiatives (CNRI) through the "Border-Safe" initiative under Grant 2030002. The Associate Editor for this paper was D. Zeng.

S. Kaza is with the Department of Computer and Information Sciences, Towson University, Towson, MD 21286 USA (e-mail: skaza@towson.edu).

J. Xu is with the Computer Information Systems Department, Bentley College, Waltham, MA 02452 USA (e-mail: jxu@bentley.edu).

B. Marshall is with the College of Business, Oregon State University, Corvallis, OR 97331 USA (e-mail: byron.marshall@bus.oregonstate.edu).

H. Chen is with the Department of Management Information Systems, University of Arizona, Tucson, AZ 85721 USA (e-mail: hchen@eller.arizona.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2008.2011695

of and explore important research questions related to cross-jurisdictional CANs.

1) What are the topological characteristics of CANs?
2) How do cross-jurisdictional data affect the topological characteristics of CANs?
3) How do CANs grow when data from multiple jurisdictions are combined?
4) How does the addition of multiple types of entities (e.g., vehicles and people) affect the topological characteristics of CANs?

In Section II, we discuss background information and previous studies. Section III presents the research testbed and design. Section IV presents the analysis of CANs using information from a single jurisdiction. Section V analyzes their characteristics in multiple jurisdictions. Section VI discusses the properties of CANs with vehicles and people, and the advantages of exploring the criminal links of vehicles. We conclude this paper in Section VII and present future directions of this research.

II. Literature Review

This section progresses from the integration of information to previous studies of criminal and other complex networks. It explains the common topological measures and discusses the evolution of networks. Previous studies on bipartite graphs are also presented.

A. Integration of Information From Multiple Sources

Cross-jurisdictional CANs contain relationships between entities like vehicles and people that are extracted from many data sources. To triangulate information about an entity, it is necessary to reconcile all the instances of the entity across data sets, which is a challenging task. Matching of entities and their relationships is a task that is hampered by problems that include [2] name differences (similar entities in different databases have different names), missing and conflicting data (incomplete data or different values in different sources), and object identification (lack of global identifiers).

We use the BorderSafe information sharing and analysis framework [3] for accessing information from multiple data sets. The key to the framework is the identification of several classes of data; the two important classes are base incident data and supplementary contact data:

1) Base incident data include information that expresses the relationships between individuals, vehicles, locations, and other such entities that are present in law enforcement incident records. For example, two individuals are related if they are partners in a crime.
2) Supplementary contact data include additional information of an annotative nature on criminal entities found in the base data. The supplementary data identify features of entities, whereas the base data express the relationships between entities. An example of supplementary data is border-crossing activity records for vehicles.

To facilitate network analysis, the base data obtained from each jurisdiction are mapped to a global schema [3]. The individuals in the base data are reconciled using first name, last name, and date of birth. The vehicles are reconciled using license plate numbers and issue authorities. The framework allows us to extract the relationships between individuals and vehicles that are amenable to CAN analysis. It also facilitates the annotation of networks with border-crossing information that can be used to identify suspicious vehicles crossing the border.

B. Complex Networks

The complex networks of individuals and other entities have traditionally been studied under the random graph theory [4]. However, later studies suggested that real-world complex networks may not be random but may be governed by certain organizing principles. This prompted the study of real-world networks. These studies have explored the topology, evolution and growth, robustness and attack tolerance, and other properties of networks. Three broad models of network topologies have emerged [4]: random graphs, small-world networks, and scale-free networks. Random graphs are networks in which any two nodes are connected with a fixed probability p; thus, edges are randomly distributed among nodes of the network. Small-world networks are not random networks and have relatively small path lengths despite their often large size [5]. In scale-free networks, the degrees (number of edges) of nodes follow a power law distribution [6]. More details on the topological characteristics of small-world and scale-free networks are discussed in the next subsection. Some of the networks that have been studied include the World Wide Web [7], [8], cellular and metabolism networks [9], and coauthorship networks [10]. These networks were found to have similar topological, evolutionary, and robustness characteristics [4]. They were found to be predominantly small world and scale free.

The structure of criminal networks has been studied using manually produced link charts [11] to depict relationships between individuals, vehicles, and locations. The topological characteristics have also been explored using social network analysis [12]–[14], shortest path algorithms [15], and manual mapping [16]. Several computerized tools like Netmap, Analyst's Notebook, and COPLINK's visualizer [17]–[19] have also been developed to support network representations of criminal activity information. To understand the topological properties of CANs and how they vary in different contexts could help investigators more efficiently perform their jobs.

C. Topological Properties

The topological properties of networks help us study the network as a whole instead of studying the individual constituents. Three concepts dominate the statistical study of the topology of networks: small world, clustering, and degree distribution [4]. These concepts have important implications in the domain of security and law enforcement.

Small World: The small-world concept is based on the fact that large networks often have small path lengths between their nodes. This is an important concept as it can depict the ease of communications within a network. Communications can range

from the spread of disease in human populations and spread of viruses on the Internet to the passage of messages and commands in a criminal network. A widely cited example of a small-world network study is the "six degrees of separation" study by psychologist S. Milgram, who concluded that there was a path of acquaintances with a typical length of about six between most pairs of people in the U.S. [20]. The small-world property is measured by the average shortest-path length that is obtained by averaging the shortest paths between all pairs of nodes in a network [4]. For instance, the average shortest-path length between two actors in a network of movie actors (225 226 nodes) was found to be 3.65 [5]. The average shortest-path length between coauthors in the MEDLINE collection (1.5 million nodes) was found to be 4.6 [10]. We will be using the actor network and the MEDLINE coauthorship network as examples to explain more concepts later in this paper. These and other examples of real-world networks show that the small-world property appears to characterize most real-world networks [4]. There has been research on the phenomenon that leads to the short path lengths in real-world networks. It has been suggested [21], [22] that shortcuts between nodes that otherwise may not be connected reduce the average path length in small-world networks. This is particularly true in social networks where people are likely to have friends with other individuals outside their immediate friend circle. The small-world property is studied in CANs because it has implications for the identification of important relationships involving suspicious vehicles and individuals.

Clustering: Cliques that represent circles of friends and acquaintances often form in social networks. For instance, authors often collaborate with the same set of people in a coauthorship network. Cliques also form in networks that do not involve people, for example, related websites on the Web often point to each other through hyperlinks. This inherent tendency to cluster is quantified by the clustering coefficient [5]. The clustering coefficient is measured by the ratio of the number of edges that exist in a network to the total number of possible edges [4]. Real-world networks tend to have relatively high clustering coefficients as compared to random graphs. The movie actors network had a clustering coefficient of 0.79 [5], and the MEDLINE coauthorship network had a coefficient of 0.066 [10], both values are several orders of magnitude higher than their random counterparts. The clustering coefficient in criminal networks points to the tendency of individuals to collaborate together and partner in crimes.

Degree Distribution: Nodes in a network have different numbers of edges connecting them. The spread of node degrees is given by a distribution function $P(k)$, which gives the probability that a randomly selected node has exactly "k" edges [4]. The distribution functions of most real-world networks follow the power law scaling with exponents ranging from 1.0 to 3.0 [4]. The movie actor network has a power law degree distribution with an exponent of 2.3 [5]. The MEDLINE coauthorship network was found to have an exponent of 1.2 [10]. Degree distributions are studied in criminal networks because high degrees of criminals may imply their leadership in the network [15]. The degrees of nodes are also used to study the growth and evolution of a network.

D. Evolving Networks

Most real-world networks, including CANs, are not static and grow due to the addition of nodes and/or links. For instance, the World Wide Web exponentially grows by the addition of new web pages, and a coauthorship network grows by the addition of collaborators. The growth leads to changes in the topological characteristics of the networks. Barabasi and Albert [4] identified two ingredients in the evolution of a scale-free network: 1) Growth: Networks continuously expand by adding new nodes; and 2) preferential attachment: New nodes preferentially attach to nodes that are already well connected, which is an effect called "rich-get-richer." The preferential attachment concept assumes that the probability that a new node will connect to an existing node i depends on the degree of the node i. The higher the degree of i, the higher the probability that new nodes will attach to it. The functional form of preferential attachment ($\prod(k)$) for a network can be measured by observing the nodes present in the network and their degrees at a particular time t. After adding new nodes $(\text{time} = t + 1)$, plotting the relative increase as a function of the earlier degree gives the $\prod(k)$ function [23]. Preferential attachment has been studied for citation, and coauthorship networks, actor network, and the Internet has been found to follow the power law distribution [23], [24]. In other cases, $\prod(k)$ may linearly grow until a point and then fall off. This usually happens at high degrees, implying that high-degree nodes are unable to attract new nodes. For instance, Newman [24] found that individuals with a large number of collaborators in a coauthorship network did not attract many new ones. These constraints on the growth of networks exist in many real-world networks including criminal networks [12], [25], [26].

The constraints on the number of links that a node can attract may be due to aging or cost [25]. Since the growth of the network may be over time, some high-degree nodes might become too old to participate in the network (e.g., actors in a movie network). It might also become too costly for a node to attach to a large number of nodes (e.g., a router in a network slows down when it has too many connections). The constraints on growth may be domain specific and have been studied in many domains. For instance, in plant–animal pollination networks, some animals cannot pollinate certain plants; hence, a link cannot be established [26]. This is an example of a cost constraint. In criminal networks, trust may restrict the growth of networks. Criminals and terrorists do not include many people in their inner trust circle [12]. In addition, disruption might restrict growth in criminal networks. Individuals may get jailed, wounded, or killed and thus not contribute to the growth. They may also "lay low" at certain times to escape the attention of authorities. These unique properties make the growth of criminal networks an interesting topic of study.

E. Bipartite Graphs

The CANs studied in this paper contain nodes like individuals and vehicles linked by police incidents. Each individual or vehicle is related to an incident, and two individuals/vehicles are linked if they are found in the same incident (more details

TABLE I
KEY STATISTICS OF TPD AND PCSD DATA

Number of records	TPD	PCSD
Incidents	2.99 million	2.28 million
Individuals	1.44 million	1.31 million
Vehicles	675,000	520,000

TABLE II
SUMMARY OF BORDER-CROSSING INFORMATION

Recorded Crossings	2.4 million
Vehicles	500,000

in Section III). These networks can be classified as bipartite graphs. Bipartite graphs contain two kinds of entities as nodes, and relationships only exist between different kinds of nodes. Many social networks like collaboration networks of movie actors, or authors, can be described as bipartite graphs [27]. Bipartite graphs are usually studied by projecting them to a unipartite graph that contains one of the entities and transitive relationships through the other entity. For instance, Watts and Strogatz [5] projected a network between actors and movies to a network between actors and actors by linking two actors who acted in the same movie. Similarly, another study of the network of directors on company boards projected a director and board network to a network of boards by linking two company boards together if they had common individuals on them [28]. We study bipartite graphs of individuals and incidents by projecting them to a network of individuals. Therefore, two individuals are linked if they are related to the same incident. The same concept is used to project networks of individuals and vehicles to networks of individuals.

III. RESEARCH TESTBED AND DESIGN

The data sets used in this paper are available to us through the Department of Homeland Security (DHS)-funded BorderSafe project. To study CANs, we used police incident reports from Tucson Police Department (TPD) and Pima County Sheriff's Department (PCSD) from 1990 to 2002. A summary of the data we used is shown in Table I.

Border-crossing information that includes the license number of vehicles with the date and time of their crossing, which is provided by Tucson Customs and Border Protection (CBP), is also included in the testbed. A summary of the border-crossing data is given in Table II.

This testbed was used to extract narcotic networks that consisted of vehicles and individuals as nodes and police incidents as edges between them. Individuals were included as nodes in the network if they were wanted, suspected, arrested, or had a warrant for arrest in a narcotics crime. Hereafter, such persons will be referred to as "suspects" in the crime. Vehicles were included as nodes in the network if they had been involved with a suspect in a narcotics crime. Two nodes were connected by an edge if they were in the same incident involving a narcotics or narcotics-related crime. An example of such a narcotic network is shown in Fig. 1. The network depicts links between individuals (circles), vehicles (rectangles), and locations (triangles) extracted from TPD and PCSD records; IT is also augmented with border-crossing information of the vehicles. The CANs used in this research do not include locations but resemble the network in Fig. 1.

To address our research questions, we divided the study into three parts. We first studied the characteristics of criminal networks in a single jurisdiction. Second, we analyzed the change in characteristics on combining data from multiple jurisdictions. In addition, we studied the characteristics of networks with multiple types of entities such as vehicles and individuals.

1) Characteristics of Criminal Networks in a Single Jurisdiction: The topological characteristics of narcotic networks extracted from TPD and PCSD were separately analyzed. Basic statistics such as size, number of links, size of giant (largest connected component), and second largest component were calculated. Small-world properties (clustering coefficient, average shortest-path length, and diameter) and scale-free properties (average degree, maximum degree, and exponent (γ) and cutoff (κ) of the degree distribution) were calculated and analyzed. This analysis aids in understanding the topological properties of narcotic networks based on the activities recorded in police records. In addition, the number of border-crossing vehicles associated with the individuals in the network was also found, which helps in identifying the criminal links of border-crossing vehicles.

2) Cross-Jurisdictional CANs: To study the topological characteristics of cross-jurisdictional CANs, we augmented the narcotics network (N) from one jurisdiction $(J1)$ with information from the second jurisdiction $(J2)$. To understand the advantages of using information from multiple agencies, the data in the second agency can be explored in two ways.

1) *Adding edges to N:* The second jurisdiction is used to identify unknown associations between the nodes of network N. The addition of incidents from $J2$ to the network in $J1$ is expected to form previously unknown associations (hidden links) among nodes in $J1$.
2) *Adding nodes and edges to N:* The second jurisdiction can be used to identify previously unknown nodes associated with the network N. The addition of individuals from $J2$ to the network in $J1$ will identify the previously unknown members of the narcotics network. The addition of nodes to N also allows us to study the phenomenon of preferential attachment in the narcotics network.

Both of the methods used to explore information in the second jurisdiction highlight the advantages of sharing data. With the addition of nodes and links from $J2$ to $J1$, we expect that the giant component of the network from $J1$ will increase in size. We also expect that the average shortest-path length of the network in $J1$ will decrease due to the addition of shortcuts between nodes.

The addition of nodes and edges from $J2$ to $J1$ can be treated as the growth of the network. Previous studies have used the preferential attachment measure to study the increase in the number of nodes over time. We use the preferential attachment measure to study the growth of the network over jurisdictions instead of over time. This sheds light on how criminals commit crimes in partnership with criminals in other jurisdictions. If the

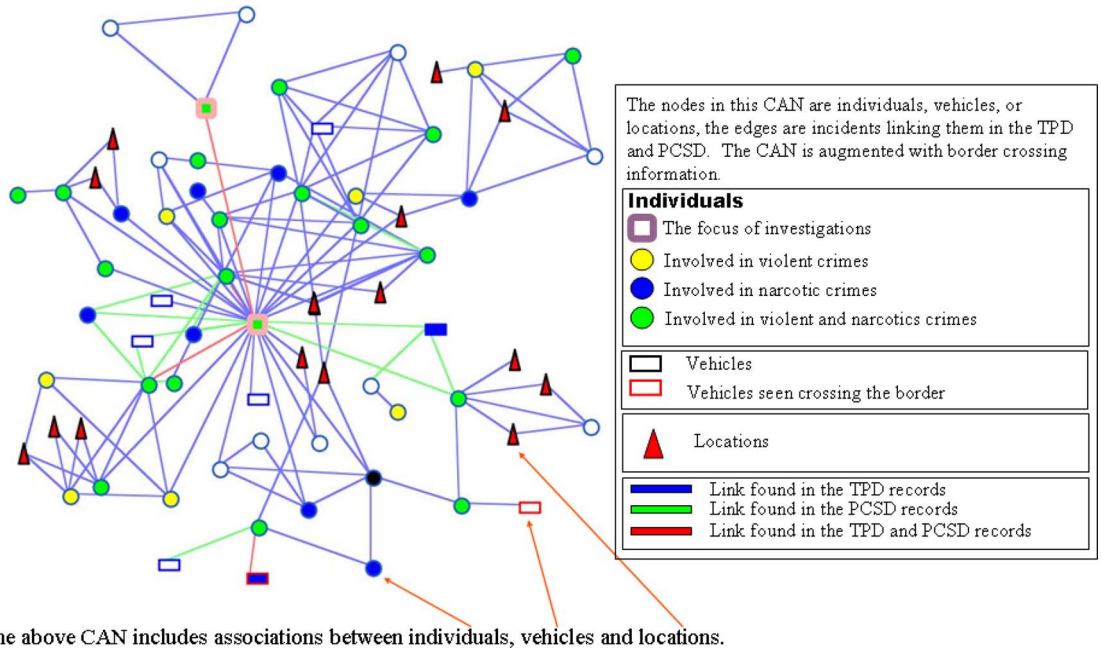

Fig. 1. Example of a narcotics CAN.

jurisdictions are geographically close or overlapping, individuals who commit more crimes in one jurisdiction will probably commit more crimes in the second jurisdiction. We expect this to be true in our data sets since Tucson, AZ (under the jurisdiction of TPD), is located within Pima County (PCSD).

3) Networks With Multiple Types of Entities: To study CANs as bipartite graphs, we define nodes as individuals or vehicles and edges as incidents that link an individual to a vehicle. This network is projected to a network of individuals by drawing an edge between two individuals who are connected through a narcotics crime to the same vehicle. This paper explores the role of vehicles in a narcotics network. We expect that the addition of vehicles will find the previously unknown links between people already present in the network. It will also help identify new members that were previously disconnected from the network. As a result, we expect the size of the giant component to increase and the average shortest-path length to decrease. Augmenting the networks with border-crossing information of vehicles can help identify the criminal links of border-crossing vehicles.

TABLE III
BASIC STATISTICS OF NARCOTICS NETWORKS

	TPD	PCSD
Nodes (individuals)	31,478	11,173
Edges	82,696	67,106
Giant component	22,393 (70%)	10,610 (94%)
2nd largest component	41	103
Associated border crossing vehicles	6,927	2,979

TABLE IV
SMALL-WORLD PROPERTIES OF NARCOTICS NETWORKS

	TPD	PCSD
Clustering Coefficient	0.39 (1.39×10^{-4})	0.53 (4.08×10^{-4})
Average Shortest Path Length (L)	5.09 (8.80)	4.62 (6.32)
Diameter	22	23

Values in parenthesis are values for a random network of the same size and average degree

IV. SINGLE JURISDICTION CRIMINAL NETWORKS

Table III presents the basic statistics of the narcotics networks extracted from TPD and PCSD's records. A giant component containing a majority of the nodes emerges from both networks. This is common in other social and affiliation networks that have been studied before [10]. The giant component in this case is a large group of individuals linked by narcotic crimes. This has important implications for social networks as the large size of the giant component coupled with short average path lengths (shown in Table IV) implies that a majority of the individuals in the network can easily be reached. In addition, we find that the second largest component is significantly smaller, suggesting that other much smaller groups of people exist in both jurisdictions. These smaller groups of criminals are likely to get connected to the giant component as time progresses.

The small-world and scale-free properties of these and other networks shown later are studied by using the giant component. The small-world properties of both networks are shown in Table IV.

The narcotics networks in both jurisdictions can be classified as small-world networks since their clustering coefficients are much higher than comparable random graphs, and they have a small average shortest-path length (L) relative to their size. *Potential applications:* The high clustering coefficient suggests that criminals show a tendency to form circles of associates who partner in crimes. According to domain experts, this is

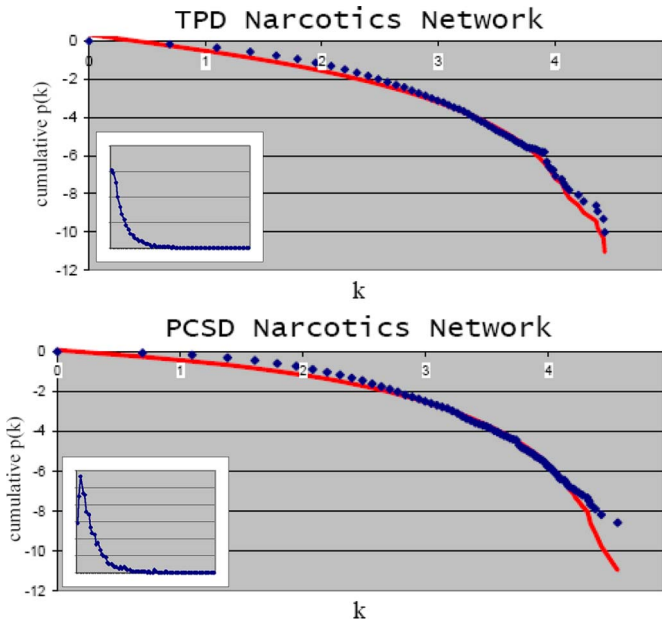

Fig. 2. Log–log plots of the cumulative degree $(p(k))$ versus the degree (k). The insets are $p(k)$ versus k. The solid line is the truncated power law curve.

TABLE V
SCALE-FREE PROPERTIES OF NARCOTICS NETWORKS

	TPD	PCSD
Average Degree, <k> (average # of partners in crime)	3.12	4.33
Maximum Degree (largest # of partners in crime)	84	96
Exponent, γ	1.3	0.85
Cutoff, κ	17.24	16.71

not unusual in narcotics networks, where individuals tend to have circles of trust that include friends and family members. This property is advantageous to law enforcement because it helps them form strong conspiracy cases against members of the group. A small L implies a faster flow of information (e.g., news of police raids) and goods (e.g., drugs) in the network. However, short paths tend to be advantageous for law enforcement too. Investigators search for associations among criminals to form a case against them. They suggest that shorter association paths between criminals generate better and higher-quality investigative leads [15].

Studies have suggested that the short path lengths in small-world networks are due to the presence of shortcuts in the network. Since the narcotics network has a short L, there must be shortcuts between people in different groups. This suggests that criminals in a narcotics network may also be committing some crimes with people outside their group.

Fig. 2 shows a plot of the degree distributions of both networks, and Table V presents their statistics.

The narcotics networks have degree distributions that follow the truncated power law, which classifies them as scale-free networks. This implies that a large number of nodes have low degrees as shown by the slow rate of decay (exponents of 0.85–1.3) at low values of k. This is expected since high degrees attract more attention from law enforcement authorities; therefore, having fewer associates is beneficial. However, it is worth pointing out that the degree of a node in these narcotics networks is also restricted by that fact that we are only considering narcotics and related crimes (to extract "pure" narcotics networks). If other common crimes like traffic citations are included, then the degrees are likely to be greater. Thus, the exponent (γ) value can be affected by the methods used for network extraction. The truncated power law distribution fits both curves better $(R^2 = 93\%)$ than the power law distribution $(R^2 = 85\%, 87\%)$. This implies that as the degree (k) increases, the probability of having k links $(p(k))$ decreases. This might indicate a cost or trust constraint to growth.

TABLE VI
TOPOLOGICAL STATISTICS ON ADDING ASSOCIATIONS (FOUND IN PCSD DATA) BETWEEN THE INDIVIDUALS IN THE TPD NARCOTICS NETWORK

Giant component	27,700 (22,393)
Edges	98,763 (82,696)
Associated border crossing vehicles	8,975 (6,927)
Clustering coefficient	0.36 (0.39)
Average Shortest Path Length (L)	8.54 (5.09)
Diameter	24 (22)
Average degree, <k>	3.56 (3.12)
Maximum degree	96 (84)
Exponent, γ	1.01 (1.3)
Cutoff, κ	16.39 (17.24)

Values in parenthesis are for the original TPD network.

V. CROSS-JURISDICTIONAL CRIMINAL NETWORKS

1) Adding Only Associations: Table VI shows the topological properties of the TPD narcotics network when it is augmented with associations found in PCSD data. No additional individuals from PCSD data were added.

In Table VI, we see that the size of the giant component in the TPD narcotics network increases. Nodes that were previously thought to be disconnected from the main network got connected. Since we only added associations, it is clear that PCSD contained associations between individuals in TPD that TPD was not aware of. The increase in the number of edges shows that previously unknown associations between existing and new nodes were added. From a total of 28 684 new relationships added, 6300 (which is a statistic not in Table VI) were between existing criminals in the TPD narcotics network. These new associations between existing people help form a stronger case against criminals.

Potential applications: The increase in the number of nodes and associations is a convincing example of the advantage of sharing data between jurisdictions.

Although we expected the average shortest path length to decrease, it increased on adding the second jurisdiction. This can be attributed to the increase in the number of nodes. Since the new nodes added did not have any associations with the existing nodes, they did not add any shortcuts to the network. However, if the giant component is not allowed to grow (no addition of nodes) and only associations between already connected nodes are added, then L decreases to 5.08. This is expected as shortcuts between the nodes are added. The average

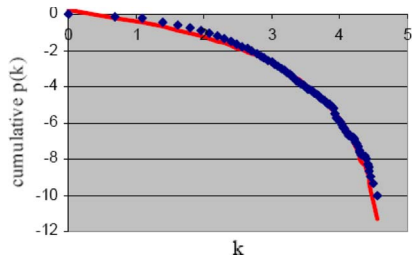

Fig. 3. Log–log plot of the cumulative degree distribution $(p(k))$ versus the degree (k) for a TPD narcotics network with PCSD links. The solid line is the truncated power law curve.

degree and the highest degree increase due to the addition of new relationships.

The number of associated border-crossing vehicles also increases. Thus, the inclusion of PCSD data provided links to more border-crossing vehicles. This will help identify more potential target vehicles at the border. Fig. 3 plots the degree distribution of the augmented TPD narcotics network. It can be seen that the network maintains a scale-free degree distribution.

2) Adding Both Nodes and Associations: To examine the results of adding both nodes and associations from the second jurisdiction, we studied the preferential attachment phenomenon. Fig. 4(a) shows the preferential attachment curve when the TPD narcotics network is augmented with both nodes and links from PCSD data. Only nodes and links connected to the nodes in the TPD network were added. Similarly, Fig. 4(b) shows the preferential attachment curve when the PCSD narcotics network is augmented with nodes and links from TPD data. Both curves lie above the solid line [in Fig. 4(a) and (b)], offering visual evidence of the presence of preferential attachment. The preferential attachment curve maintains linearity for small values of k but breaks down for higher degrees. This can be attributed to the nature of the networks being studied. Criminals may not prefer to be related to a large number of individuals for the risk of drawing attention. Thus, the cost of acquiring more links is high; this might prevent a node with a large number of links to acquire more. In addition to the cost effect, there may be various other reasons that may encourage or discourage the formation of links between criminals. These can also be used for link prediction and are explored in detail in [29]. Additionally, external influences like law enforcement limit the number of crimes that an individual can commit. Third, the higher degree nodes may not attract more nodes as they may not be committing crimes in the second jurisdiction, or conversely, the lower degree nodes attract more nodes as they commit more crimes in the second jurisdiction. This is possible if one jurisdiction had incomplete information on some of the criminals in the network, and therefore, they ended up with lower degrees. Combining the information from the second jurisdiction revealed more crimes and increased their degrees.

VI. Multiple Entity Criminal Networks

To study multiple entity networks, we added individuals who were linked to the existing nodes through vehicles to the TPD narcotics network (as described in Section III-C). Table VII shows the change in topological properties of the TPD and PCSD narcotics networks on adding transitive links through vehicles.

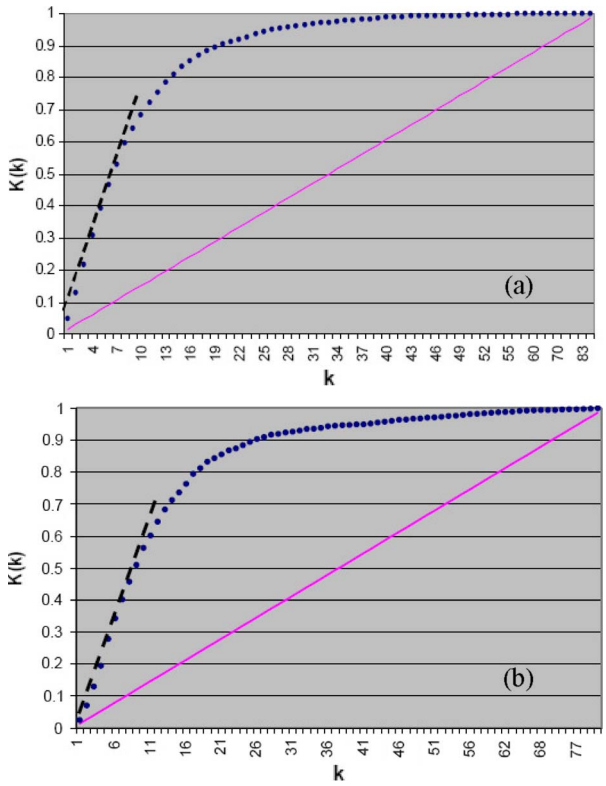

Fig. 4. Curve shows preferential attachment when the narcotics network in (a) TPD is augmented with data from PCSD and (b) PCSD is augmented with data from TPD. The dashed line above the curve shows a linear preferential attachment growth, and the solid line shows the state of no preferential attachment.

TABLE VII
Topological Properties of TPD and PCSD Narcotics Networks on Adding Links Through Vehicles

	TPD	PCSD
Giant component	26,797 (22,393)	13,560 (10,610)
Edges	81,263 (70,079)	49,853 (46,004)
Associated border crossing vehicles	7,985 (6,927)	3,075 (2,979)
Clustering coefficient	0.37 (0.39)	0.43 (0.53)
Average Shortest Path Length (L)	6.05 (5.09)	4.65 (4.62)
Diameter	24 (22)	26 (23)
Average degree, <k>	3.03 (3.12)	3.67 (4.33)
Maximum degree	94 (84)	96 (96)
Exponent, γ	1.1 (1.3)	1.1 (0.85)
Cutoff, κ	19.04 (17.24)	19.34 (16.71)

Values in parenthesis are for the single entity (individuals only) networks.

Linking individuals through vehicles created links between people who were previously not known to be associated. This increased the size of the giant component of the networks.

Potential applications: The additional association to border-crossing vehicles implies that the criminal history of these vehicles can be extracted to aid in the identification of suspect vehicles at the border (details on a method to use criminal

histories to identify such vehicles can be found in [30]). Since not all potentially suspect border-crossing vehicles have criminal histories recorded in local law enforcement databases, finding such additional associations can be beneficial. The slight increase in the average shortest-path length may be attributed to the fact that the networks now contain individuals who do not have many links to existing nodes but are associated only through vehicles. These individuals do not add any shortcuts to the existing networks and serve to increase the path length. The decrease in the exponent (γ) may be due to the increase in the number of low-degree nodes. Overall, the addition of vehicles provides links to individuals who are not directly associated with criminals. This aids law enforcement and also helps identify the criminal links of a vehicle. The amount of criminal activity of a vehicle in border-area jurisdictions can be used to identify suspect vehicles at the border. A vehicle with high criminal activity can be more thoroughly examined by customs and border-protection agents.

VII. Conclusion and Future Directions

CANs extracted from multiple law-enforcement- and transportation-related data sources can be used to aid in border protection and transportation security. This paper has focused on the topological properties of CANs in a cross-jurisdictional context. The role of vehicles in narcotics networks was also studied. Narcotics networks were found to be small world in nature with short path lengths and scale-free degree distributions. These topological properties have important implications for law enforcement and, hence, transportation security. It was found that a single jurisdiction may contain incomplete information on criminals, and cross-jurisdictional data provide an increased number of high-quality investigative leads. The inclusion of vehicular data in CANs had clear advantages. Vehicles provided new investigative leads that can be used to target individuals and vehicles that might pose a threat to the security of the border and transportation infrastructure.

In the future, the robustness and attack tolerances of criminal networks will be studied. Robustness analysis can be used to identify the best attack strategies to break down the networks. The topological characteristics of other networks like car-theft rings, gang networks, and networks with locations can be studied to understand the difference between these criminal networks.

Acknowledgment

The authors would like to thank the BorderSafe project partners: Tucson Police Department, Pima County Sheriff's Department, Tucson Customs and Border Protection, Automated Regional Justice Information Systems (ARJIS), San Diego Super Computer Center (SDSC), Department of Homeland Security, and Corporation for National Research Initiatives (CNRI); H. Atabakhsh and H. Gowda from the AI Lab, University of Arizona, and T. Petersen from the Tucson Police Department for their contributions to the research presented in this paper; and M. Patton and E. Skidmore of the Hoffman Ecommerce Lab, University of Arizona, for providing high-end computing support.

References

[1] *National Strategy for Homeland Security*, 2002, U.S. Office of Homeland Security.
[2] I.-M. A. Chen and D. Rotem, "Integrating information from multiple independently developed data sources," in *Proc. 7th Int. Conf. Inf. Knowl. Manage.*, Bethesda, MD, 1998, pp. 242–250.
[3] B. Marshall, S. Kaza, J. Xu, H. Atabakhsh, T. Petersen, C. Violette, and H. Chen, "Cross-jurisdictional criminal activity networks to support border and transportation security," in *Proc. 7th Int. IEEE Conf. Intell. Transp. Syst.*, Washington, DC, 2004, pp. 100–105.
[4] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," *Rev. Mod. Phys.*, vol. 74, no. 1, pp. 47–97, Jan. 2002.
[5] D. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.
[6] A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Sci.*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.
[7] R. Albert, H. Jeong, and A.-L. Barabasi, "Diameter of the world-wide web," *Nature*, vol. 401, pp. 130–131, 1999.
[8] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal, "The web as a graph," in *Proc. 19th ACM SIGACT-SIGMOD-AIGART Symp. Principles Database Syst.*, 2000, pp. 1–10.
[9] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, Oct. 2000.
[10] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Nat. Acad. Sci.*, vol. 98, no. 2, pp. 404–409, Jan. 2001.
[11] W. R. Harper and D. H. Harris, "The application of link analysis to police intelligence," *Hum. Factors*, vol. 17, no. 2, pp. 157–164, 1975.
[12] P. Klerks, "The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands," *Connections*, vol. 24, pp. 53–65, 2001.
[13] M. K. Sparrow, "The application of network analysis to criminal intelligence: An assessment of the prospects," *Soc. Netw.*, vol. 13, no. 3, pp. 251–274, Sep. 1991.
[14] J. Xu, B. Marshall, S. Kaza, and H. Chen, "Analyzing and visualizing criminal network dynamics: A case study," in *Proc. ISI*, Tucson, AZ, 2004, pp. 359–377.
[15] J. Xu and H. Chen, "Fighting organized crime: Using shortest-path algorithms to identify associations in criminal networks," *Decis. Support Syst.*, vol. 38, no. 3, pp. 473–487, 2004.
[16] V. E. Krebs, "Mapping networks of terrorist cells," *Connections*, vol. 24, no. 3, pp. 43–52, 2002.
[17] *I2 Investigative Analysis Software*, vol. 2004.
[18] *COPLINK from Knowledge Computing Corp.*, vol. 2004.
[19] E. Chabrow, "Tracking the terrorists: Investigative skills and technology are being used to hunt terrorism's supporters," *Inf. Week*, Jan. 14, 2002.
[20] M. Kochen, *The Small World*. Norwood, NJ: Ablex, 1989.
[21] T. Nishikawa, A. Motter, Y.-C. Lai, and F. Hoppensteadt, "Smallest small-world network," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 66, no. 4, p. 046 139, Oct. 2002.
[22] D. Watts, *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton, NJ: Princeton Univ., 1999.
[23] H. Jeong, Z. Neda, and A.-L. Barabasi, "Measuring preferential attachment for evolving networks," *Europhys. Lett.*, vol. 61, pp. 567–572, 2003.
[24] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 2, p. 025 102, Aug. 2001.
[25] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, "Classes of small-world networks," *Proc. Nat. Acad. Sci.*, vol. 97, no. 21, pp. 11 149–11 152, Oct. 2000.
[26] P. Jordano, J. Bascompte, and J. M. Olesen, "Invariant properties in coevolutionary networks of plant–animal interactions," *Ecol. Lett.*, vol. 6, no. 1, pp. 69–81, Jan. 2003.
[27] M. E. J. Newman, S. H. Strogatz, and D. Watts, "Random graphs with arbitrary degree distributions and their applications," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 2, p. 026 118, Aug. 2001.
[28] G. F. Davis, M. Yoo, and W. E. Baker, "The small world of the American corporate elite, 1982–2001," *Strategic Org.*, vol. 1, no. 3, pp. 301–326, 2003.
[29] S. Kaza, Y. Wang, and H. Chen, "Suspect vehicle identification for border safety with modified mutual information," in *Proc. Intell. Security Informat.*, 2006, vol. 3975, pp. 308–318.
[30] S. Kaza, Y. Wang, and H. Chen, "Enhancing border security: Mutual information analysis to identify suspect vehicles," *Decis. Support Syst.*, vol. 43, no. 1, pp. 199–210, Feb. 2007.

Siddharth Kaza received the Ph.D. degree in management information systems from the University of Arizona, Tucson, in 2008.

He is currently an Assistant Professor with the Department of Computer and Information Sciences, Towson University, Towson, MD. His research interests include social network analysis, data mining and knowledge discovery, decision support systems, and security informatics.

Jennifer Xu received the Ph.D. degree in management information systems from the University of Arizona, Tucson, in 2005.

She is currently an Assistant Professor with the Computer Information Systems Department, Bentley College, Waltham, MA. Her research interests include knowledge discovery and data mining, knowledge management, social network analysis, information visualization, human–computer interaction, and electronic commerce.

Byron Marshall (M'05) received the Ph.D. degree in management information systems from the University of Arizona, Tucson, in 2005.

He is currently an Assistant Professor of information management with Oregon State University, Corvallis. His research interests emphasize the reuse of organizational data in informal node-link knowledge representations to support analysis tasks. His previous work includes applications in bioinformatics, business intelligence, digital library, law enforcement, and education.

Hsinchun Chen (M'92–SM'04–F'06) received the B.S. degree from National Chiao Tung University, Hsinchu, Taiwan, the M.B.A. degree from the State University of New York (SUNY), Buffalo, and the Ph.D. degree in information systems from New York University, New York, NY.

He is currently a McClelland Professor of management information systems with the University of Arizona (UA), Tucson. He has served as an Advisor for major NSF, DOJ, NLM, DOD, DHS, and other international research programs in digital library, digital government, medical informatics, and national security research. He is the Founding Director of the UA Artificial Intelligence Laboratory and the Hoffman Ecommerce Laboratory. He is author/editor of 18 books, 17 book chapters, 150 SCI journal articles, and 110 refereed conference articles covering Web computing, search engines, digital library, intelligence analysis, biomedical informatics, data/text/web mining, and knowledge management.

Dr. Chen is a Fellow of the American Association for the Advancement of Science (AAAS). He serves on ten editorial boards, including the *ACM Transactions on Information Systems*, the *ACM Journal on Educational Resources in Computing*, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, the *Journal of the American Society for Information Science and Technology*, *Decision Support Systems*, and the *International Journal on Digital Libraries*. He was awarded the Andersen Consulting Professor of the Year in 1999.