Comput Math Organ Theory (2010) 16: 89–111
DOI 10.1007/s10588-009-9060-8

Hiding in plain sight: criminal network analysis

Christopher E. Hutchins · Marge Benham-Hutchins

Published online: 26 November 2009

Abstract The United States is faced with an increasingly complex criminal enterprise. Advances in technology, communications, transport, and economies enable a highly adaptive criminal element to hide in plain site. These advances provide criminal organizations with the same global boundaries and opportunities as legitimate organizations.

As boundaries expand the data to be analyzed by law enforcement mounts at a geometrically astounding rate. In response, the nature of law enforcement intelligence analysis must evolve to cope with both the amount and complexity of the data. This requires new and adaptive methods of analysis.

Researchers have found that the principles of network analysis can be applied to the analysis of terrorist and criminal organizations. This paper examines the combination of measures historically employed by intelligence analysts and network analysis software and methodologies to quantitatively and qualitatively examine criminal organizations.

Keywords Social network analysis · Dynamic network analysis · Criminal network analysis · Law enforcement · ORA · HIDTASIS

Abbreviations

SNA	Social Network Analysis;
DNA	Dynamic Network Analysis;

C.E. Hutchins
North Texas High Intensity Drug Trafficking Area (NTHIDTA), Irving, TX, USA
e-mail: chrisehutch@yahoo.com

M. Benham-Hutchins (✉)
School of Nursing, Bouvé College of Health Sciences, Northeastern University, 360 Huntington Ave, Boston, MA 02115, USA
e-mail: m.benhamhutchins@neu.edu

CASOS	Center for Computational Analysis of Social and Organizational Systems;
HIDTASIS	High intensity drug trafficking area strategic information system;
LEA	Law enforcement agency

1 Introduction

The United States is faced with an increasingly complex criminal enterprise. Advances in technology, communications, transport, and economies enable a highly adaptive criminal element to hide in plain sight. These advances provide criminal organizations with the same global boundaries and opportunities as legitimate organizations. As boundaries expand the data to be analyzed by law enforcement mounts at a geometrically astounding rate.

In response, the nature of law enforcement intelligence analysis must evolve to cope with both the amount and complexity of the data. This requires new and adaptive methods of analysis. This paper presents three sample cases which illustrate the integration of a custom high intensity drug trafficking area strategic information system (HIDTASIS) application used for data mining and extraction and dynamic network analysis (DNA) software applications developed by researchers at the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University.

Historically criminal network analysis has been a process requiring human analysts to sort through numerous data resources and manually draw diagrams and compile data; this time consuming process has rapidly become obsolete in the face of a sophisticated and technologically advanced criminal enterprise. Mass communication media and transportation infrastructure have served to decouple crime with location. Simple "who, what, where" investigations rapidly escalate geometrically with each connection. The artifacts of these communications and travels result in masses of data which must be examined and salient facts selected to focus and advance the investigation. Researchers have found that the principles of social network analysis (SNA) can be applied to the analysis of terrorist and criminal organizations (Stewart 2001; Xu and Chen 2005). Integrating the historical measures and methods used by analysts, sophisticated network analysis software and SNA methodologies provide the means to mine and study large amounts of data, uncovering both the relationship between members and the physical structure of the criminal network. This new knowledge assists with criminal investigation and provides guidance in the disruption of criminal networks.

2 The new order

Traditionally, crime had been thought to be a local behavior and local means were employed to combat it. In this shrinking world, with personal mobility at new high levels and worldwide communication tools in the hands of virtually everyone, law enforcement agencies (LEAs) are often finding their jobs involve criminal networks. These networks may be only loosely hierarchical, and frequently

formed on an ad hoc basis to suit a particular short term goal supporting local and remote criminals. For example, drug traffickers are investigated on local, state, and federal levels. The High Intensity Drug Trafficking Area (HIDTA) program was developed to create coalitions of LEAs for this purpose. Congress established this program under the Office of National Drug Control Policy as an adaptation to the organized transshipment of illicit drugs from foreign countries, recognizing these activities were from networked sources. McGloin (2005) recognizes network analysis as a technique that identifies patterns though which certain gangs "nominate themselves" for focus of investigation. Ressler (2006) points out that "networks are ubiquitous, with underlying order and simple laws" (p. 1) and encourages a network analysis approach to anti-terrorism efforts based on terrorist networks typically being non-hierarchical and spanning countries and continents. Network measures have been used to identify criminal networks (Xu and Chen 2005; Xu and Hsinchun 2005). Williams (2001) develops the case that even organized crime operates through fluid networks rather than through more formal hierarchies. He goes on to describe network features and individual roles in criminal networks that we see highly reflected in anecdotal evidence in drug trafficking cases.

As the knowledge and techniques of "Third Generation" network analyses (Klerks 2001) and tools filter into the law enforcement arena, they promise to change the face of modern criminal network analysis. Law enforcement analysts now have the ability to use advanced software to examine criminal network data and uncover patterns and linkages between individuals, a specific crime, geographic location and/or resource such as a car or gun (Innes et al. 2005). This paper presents an experiential application of "third generation" network analysis techniques and tools accomplished by integrating DNA software and the information available through a customized, web enabled database application.

3 Background

In 2000, based on the limitations and cost of existing strategic management systems, the decision was made to develop a custom case management system that would overcome the limitations of available systems and support data transfer from existing databases. Within 6 months the first version of the HIDTASIS was introduced. During the ensuing years, development of the application continued and it is currently a stable, mature web application running on a universal platform (JAVA) supported by an open source database (MySQL). Figure 1 illustrates the HIDTASIS network infrastructure. There is no hidden code or proprietary licenses to inhibit use of the HIDTASIS software by any LEA in the country.

The original HIDTASIS software design incorporated the "case-centric" design similar to its predecessors. The analytic process was assumed to start with the definition of a case which became the governing identification to which all records would be related. Limitations to this design soon became apparent with the realization that data becomes available and important to retain irrespective of the confines of a specific case. In fact, during the first 10 months of data entry, nearly 80 percent of the data entered used the term "pending" for the case identification. The case-centric design also limited the ability of analysts to link independent data observations to other

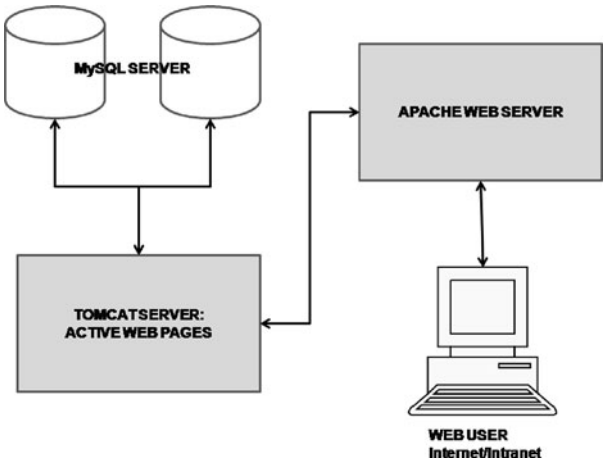


Fig. 1 HIDTASIS Architecture

data elements. To overcome this limitation, the system was redesigned to allow all forms of data relations. The viability of this modification was soon proven in the field by gang intelligence units in a major U.S. city using the system to track gang members independent of the specification of a particular case.

HIDTASIS is an observation oriented application rather than a monolithic case structured system. Users are free to develop their own input style of data entry or adapt to specific data entry approaches. They can enter all the data regarding a related set of persons, things, events, etc., or enter lists of items and link them afterward.

The move away from the case-centric design led to the move toward dynamic, multi-modal network analysis since investigators linked persons both directly and indirectly through other tracked resources. Although the design changes opened the use of data to better analysis, the presentation of data remained in static tables and reports spread over numerous pages with no succinct visual rendering. Organizational Risk Analyzer (ORA) and Automap, network analysis software developed by researchers at Carnegie Mellon University, were identified as compatible with the HIDTASIS software (Carley and Reminga 2004; Diesner and Carley 2004, 2005, 2008). Early work with these programs revealed a highly flexible, efficient means of leveraging electronic data of all sorts.

The tools that comprise the ORA and Automap programs enhance the analytical capabilities of the HIDTASIS in different ways. For example, since the HIDTASIS was designed to link entities, it was a logical extension to export table data through a programmed process for persons, cases and predefined criminal organizations. In addition, extraction of network data from text is accomplished by Automap; this requires use of a semi-automatic process developed to extract information from text files and other database fields. Importing the extracted data into the ORA program allows analysis based on network structure and measures.

Multimodal analysis allows linkages and analysis based on a task, resource, knowledge and/or location. The math and models that make up DNA provide insight into real world problems, since the modern criminal network has undergone specialization, often involves transportation of illicit goods, and engages in sophisticated

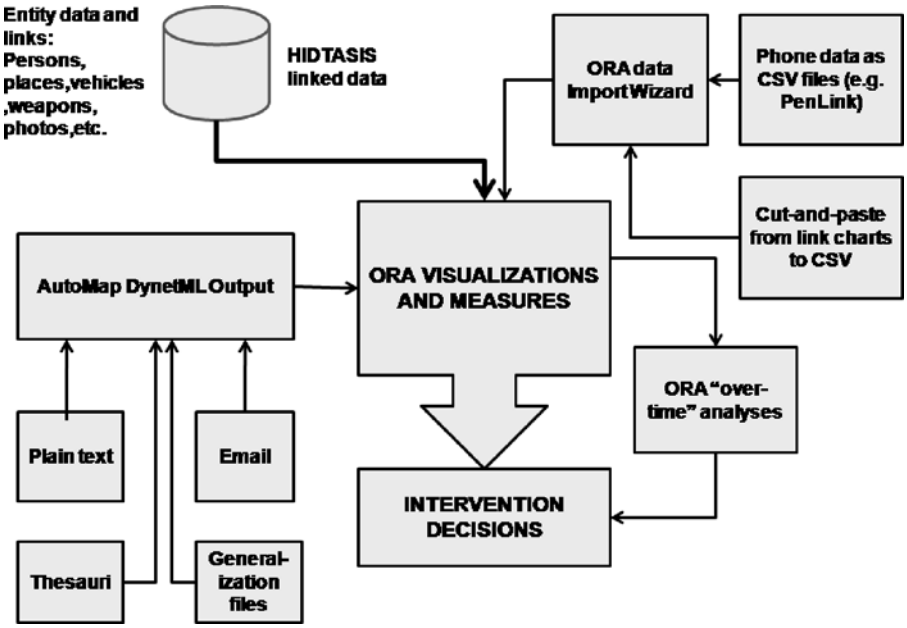


Fig. 2 Information flow and analysis process diagram

fiscal transactions. And, notably, these computations and models provide scalable results for the networks found in law enforcement efforts (Ozgul et al. 2007). Changing the HIDTASIS from a case centric design to one that allows free linkage of independent data elements was an important step in moving toward dynamic, multimodal criminal network analysis.

4 The data process: availability and manipulations

4.1 Importing data

Incorporating data from all paper and electronic law enforcement records provides the information necessary to optimally use DNA. Perhaps the most common barrier to the use of a new software option is the ability to use information and data available in existing databases. There are two primary types of available data: unstructured and structured data. These data are available from multiple paper and electronic sources and must be prepared for analysis. This is accomplished by pairing the HIDTASIS with ORA and/or Automap. The import wizard and integrated XML scripting allow an analyst with moderate computer skills to bridge this gap. Importing data for network analysis is a semi-automatic procedure (Fig. 2).

4.2 Unstructured data

Unstructured paper-based textual data makes up a significant percentage of law enforcement records. The growth of technology has provided tools for converting paper

files to electronic with rapid, high accuracy scanners and optical character recognition programs. Precisions greater than 98 percent are common for software through self-correcting algorithms in both the creation of the text image and its recognition. Unstructured electronic images can be organized using image matching software. While still in its infancy, it may be used for facial recognition and automatic photo indexing.

4.2.1 Automap

To prepare existing or converted electronic unstructured data (text files) for use in ORA, it is first categorized using Automap. Automap has been developed to extract from a single electronic text or from multiple electronic documents. The basic premise is that when people write sentences, the structure of word combinations contains information that by position and/or part-of-speech detection indicates relations between subjects. The software examines the structure and content of the text with language standards and specific requirements set by the user. Results are available in a specialized XML file format (DynetML) that can be imported directly into ORA (Carley et al. 2007).

Of the preprocessing operations built into the software, two are the most critical with respect to outcome of the extraction. They are the Named Entity (NE) Generalization Thesaurus and the Meta-Matrix Thesaurus. The NE thesauri are the backbone of networks because they are the root target of extraction. The Meta-Matrix Thesaurus defines what entity sets the NE are placed in and what additional information will be extracted as attributes for nodes.

The NE extraction option in Automap is a series-capitalization extraction only. This is disappointing for two reasons. The first is that it fails at a high rate when it encounters Hispanic and Asian names. These naming schemes are much more complex than European and American practices. Many of the names consist of more than 2 elements and multi-element names may have one or more qualifiers, such as “de la Paz” and others. The second issue is the sole choice of considering only capitalization for selection. Many legacy databases did not provide for upper and lower case entry and numerous agencies use all uppercase as a standard practice even though the systems are capable of mixed case. Additionally, it has become common practice in electronic mail to disregard capitalization and in chat practice, uppercase is considered as “yelling.” Since we intend to submit all these types of data to Automap, it seems an important deficiency.

To compensate for this deficiency, an Open Source alternative was identified to generate NE thesauri (http://nlp.stanford.edu/software/CRF-NER.shtml). Stanford Named Entity Recognizer is a java application that can be run to produce a tagged text file of NE which can be subsequently read into a database with ORA-friendly meta-entity over-coding via our internal Tomcat server. If this seems complex, be assured it is, but also know that it is easily understood after a reasonable learning curve. It is also a script capable program.

Business and government researchers are pursuing NE extraction techniques. The ability to reduce millions of documents in thousands of locations to actionable intelligence is a goal shared by all knowledge consumers, regardless of their discipline. We

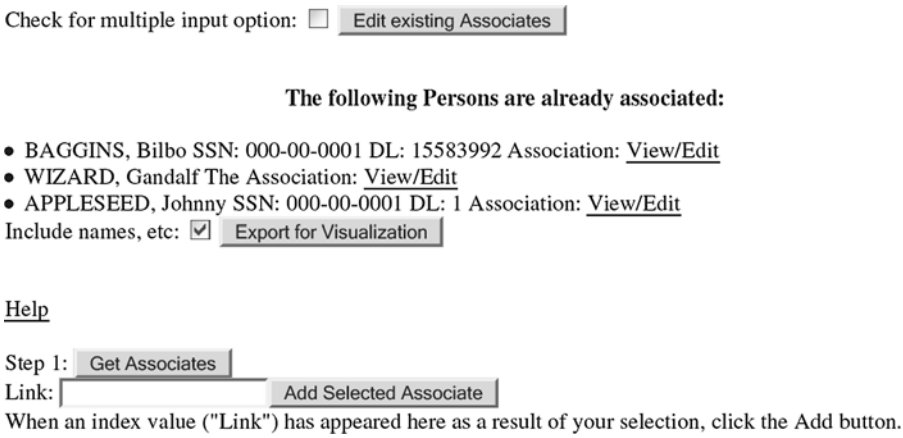


Fig. 3 Export option on HIDTASIS associate's control panel

expect order of magnitude improvements in this technology in the next 5–10 years and expect to leverage any and all we can find that are compatible with our data. The researchers at CASOS are well aware of these facts and have a committed philosophy toward maintaining these tools as a suite capable of importing and exporting information easily.

4.3 Structured data

Although it would be beneficial to have all structured data in the HIDTASIS database, thus forming a single source for relational data, the process of entering the data currently on-hand is impractical for most of our other structured sources. The ORA data import wizard handles the burden for all the miscellaneous comma-separated-value files exported from other analyst software programs. Since these files are often the only source of electronic data available from the static analysis programs in general use, they are a major source of data. The intuitive ORA interface supports import of more than 100,000 records in just seconds. The import requires selecting which columns will be treated as entities and then defining the resulting networks. Optionally, it is easy to define attributes for entities, even combining column data to create meta-properties.

Structured data available through HIDTASIS is accessed and transferred to the ORA program through the use of the HIDTASIS program tools combined with the data import wizard from ORA. Because of the structured nature of databases and the link capability of HIDTASIS in particular, we are able to export/import networks to ORA quite simply. The easiest method using HIDTASIS is to select a person or organization in the database and click on the visualize button available on the HIDTASIS dashboard (Fig. 3). This action will access corresponding records and export a four generation data array of nodes and relations of persons, vehicles, phones, addresses, etc. The HIDTASIS server accomplishes this through a snowballing discovery process that traces each pathway to its endpoint (the endpoint is either generation four or no further data discovered whichever comes first).

The output is appended to tables accessible to ORA as a new data set or appended to an existing data set. This importation is manual but simple to perform and quite fast. Multiple entity sets and graphs may result from the HIDTASIS operation, an entity set for each type of data linked and at least multimode graphs for each data type pairing.

Attribute data for any given node is not included with this export method and is not necessary for the computation of network measures. If attribute data is desired, a scripted import is required which is initiated from the ORA data import wizard rather than HIDTASIS. Attribute data has the potential to enhance the ORA visualization because the nodes can be colored, sized, or grouped based on common characteristics.

4.3.1 Data transformation: an evolving process

In some respects this description of the process may seem to start in the middle and work back to the beginning. To a large degree that is how the process was developed. One of the guiding principles used in evaluation and development of HIDTASIS is the ability to maximize use of existing data in its current form, providing data mutability that is as transparent to the user as possible. Current efforts are focused on methods that enhance the ability of the analyst to easily organize and incorporate unstructured data for criminal network analysis.

Although structured data is more accessible it is important to be careful during evaluation to avoid misrepresentation or misinterpretation. Historically, database tables are often linear and an identifying column is only related to attributes in its own row. It takes more thought in table development to be able to provide edge linkages. Most relational databases were not designed to provide person-to-person relational data or, in fact, any entity-to-entity relations (Coles 2001). Database schema that handle multidimensional data are carefully constructed to employ table linkages that accomplish the links that users "see."

Another process that depends on existing data is the discovery of inferred links. Inferred links are links where no direct connection is evident but entities that share links to another entity may also be linked to each other. Using such links in analysis will result in hypothetical network states. Probabilistic mathematical studies are the focus of current research and the current version of ORA includes probabilistic tools such as the ability to fold a network matrix such that first generation inferences can be clearly visualized.

4.3.2 Scripting

Preparing data for import into ORA can be facilitated by the use of XML scripting. These scripts, written in human readable code, read data from other files for use in constructing the networks for ORA to analyze. Once the mechanics of the script versus the outputs is understood; a highly tailored analysis can be set up and reproduced, including extracting longitudinal data in a single operation. This is useful in time related sequence analysis and network development studies. In the current scripting model the data query can be different in each loop with the graph and node data outputs fixed. This gives the analyst the opportunity to ask very different questions of the

Table 1 Triangulation analysis process

- Load a **Meta-Network** in **ORA**.
- Click the **View Charts** button in the right panel.
- From the **Bar Chart** panel, select measures from the drop down menu above the displayed chart: Set the ranked number of entities to show in the bar chart (use a low number, such as 5).
- Open the **Visualizer** from the ORA navigation window (select a visualization which will include the entities from the bar chart).
- Select **Sphere of Influence** from the **Tools** menu.
- In the **search/filter textbox**, enter enough information to find the entity of interest from the Bar Chart. Click the checkbox next to that entity in the listing. If visualizing for **Clique Count** ranking, select a radius of 1 in the radius box at the top of the dialog. For **Centrality** measures and **Potential Boundary Spanner** measure, select a radius of at least 2.
- **Animate** the resultant visualization. If visualizing for Clique Count, it's useful to hide pendant nodes, since they have only 2 members are not defined as cliques (minimum 3).
- On the **File** menu, select **Visualize** in 3D.
- **Repeat** this process for the top ranked one or two entities for the network in each measure where there are large differences between the top ranked and bottom ranked measure values or there are multi-modal distributions, normal distribution, or positive-sloped distributions indicated by the Histogram for that measure.

data for the same or similar node and graph structures. In addition, analysts can share and re-use scripts with minor changes. Automap can be scripted very easily to select multiple file inputs from a single directory and to use prepared lists and thesauri to perform standardized analyses. This output is manually imported into ORA.

5 Dynamic network analysis

Once the data is imported into the ORA program the next step is to process the information and proceed with analysis of the criminal network. The program supports both qualitative (visual) and quantitative (mathematical) analysis of the network data. Although there is a tendency to separate the analysis of the network visually and the analysis of the network mathematically a triangulation process outlined here provides a mixed method approach that builds on the strengths of both types of analysis by preserving both the visual pictures and the math. Table 1 provides an overview process for examining a criminal network using dynamic network analysis tools.

5.1 Applying network analysis

Data from three networks is presented to illustrate findings identified through organizational risk analysis. The source of Network 1 data is street gang data stored in HIDTASIS. The agents in this data are street gang members that have themselves claimed membership, been identified by previously proven members, or been involved with numerous events associated with the gang. This export produced 215 agent nodes, 217 location nodes, and 23 resource nodes. In addition, the network links to 80 organizations and 834 events. The event entries in this database are comprised of each police contact report associated with any of the agent entities. The records span more than six years and the data was collected by a large metropolitan police department. This produces an extraordinarily rich data set for analysis.

Network 2 is a network formed by analysis of 147,629 phone calls acquired by subpoenas issued for principals involved in a major drug investigation. Although it is uni-modal, certain observations can be made by analysis of the telephone subscriber information which shows the surrogacy of the phone number entity to persons or businesses or government agencies. This network has been reduced to 933 agent entities in its current form. Commonly called toll analysis, this is a typical data set, both in size and scope, encountered in these investigations. Most analysts can create a chart of the phone connections between agents but then must manually review the chart to discover what links are important, a process which may take weeks and is extremely subject to bias. The principal metric available in these charting methods is call frequency which may mean little or nothing to the business logic of the drug traffickers behind the calls. Applying social network measures to the data set reduces time spent on unproductive connections and quickly focuses resources toward viable leads.

Network 3 is a multi-modal network of agents, resources, locations, events, knowledge, roles, and task entities. This network was manually constructed from cell phone data extractions and information identified from the physical evidence present at the scene of a highway traffic stop that found a large sum of money concealed in a vehicle. Drug trafficking networks are, by intent, covert networks. Traffic stops which locate contraband are being emphasized as intelligence sources because this is a point at which the covert network becomes visible. By maximizing the gathering and interpretation of evidence found at the scene and exploiting it leads to discovery of the network. In this instance, 259 agents were identified, with the involvement of 16 locations, 2 organizations, 7 events, 5 resource entities, and 325 knowledge components. From the data, analysts linked the two principals to a task-role network designed to identify intelligence gaps and further the discovery of the whole network. Agencies at the source region and the destination regions were provided with the data.

In the following case analyses all measures, tables and graphs were produced using the ORA[1] software. Selection of reports is discussed in the Appendix, "Using Charts and Reports."

6 Network 1

Our analysis goal for Network 1 is monitoring and observation. This network is a moderate size urban gang and its membership, territory, and affiliations are of interest in the event of violence or organized criminal activity. Starting with the visualizer option, an overview of the network structure can be reviewed. Selecting a subgroup identification method is a common component of network visualization. For example, the analyst may select one or more of the node-sets and begin looking for patterns. This may be necessary because the nodes and edges of the visible network are too

[1]The Organizational Risk Analyzer, Copyright © Kathleen M. Carley, Center for Computational Analysis of Social and Organizational Systems (CASOS), Institute for Software Research International (ISRI), School of Computer Science (SCS), Carnegie Mellon University (CMU), 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, http://www.casos.cs.cmu.edu.

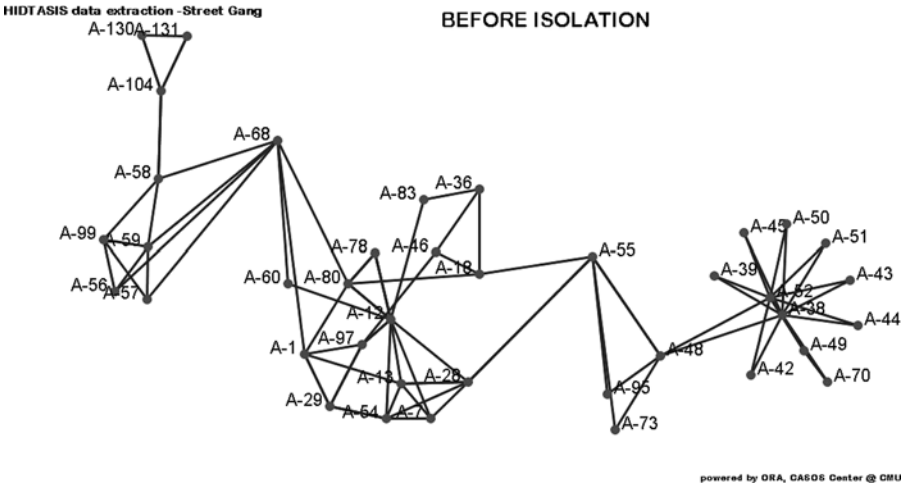


Fig. 4 Gang subgroup network structure

dense to determine individual characteristics. There are also a number of tools to trim down the view, such as isolate removal and dropping pendants. There are multiple ways to accomplish this, some only affect the viewed nodes when selected in the visualizer, not the original file, but other methods, such as the transform option, change the selected network data. It is important to be aware of the potential for data loss when selecting one of these options.

Isolated nodes were removed from Network 1 which reduced this to 171 nodes. Hiding pendant nodes further reduced the network to 100 nodes. We colored nodes by component and that identified three components, the largest of which is presented in Fig. 4, a final node count of 73 with 273 total links. This nodeset is all those nodes that are connected to more than 1 entity in the overall network. This in effect, reduces the network to a bounded set. This process is employed for all but the simplest networks. When the network is reduced to a manageable set, the analyst can request standard quantitative measures for the visible network. The results of these calculations can be utilized to size or color nodes to highlight nodes with specific characteristics.

Figure 5 provides a list of the top ranked members (nodes: A-12, A-80, etc.) of the gang depicted in Fig. 4. Although they are highly ranked, they may not be the best targets for intervention by law enforcement (Klerks 2001). For disruption purposes, greater effect may be achieved by selecting from the table of Connects Groups (Potential Boundary Spanners) in this report (Fig. 6).

These are the "brokers" in our network who are potentially the keys to the survival of the joined nodes economically or socially (Boissevain 1974). Note that the targeted nodes, A-55 and A-68, are in the middle of this list. As illustrated in Fig. 7, removal of these nodes changed the network structure.

Two network properties of key members are *Clique Count* and *Betweenness Centrality*. A clique, where each node is connected to every other node, is an obvious matter of concern in interdependence and subgroup analysis. *Betweenness Centrality* is a measure of importance as it identifies nodes which provide a communication

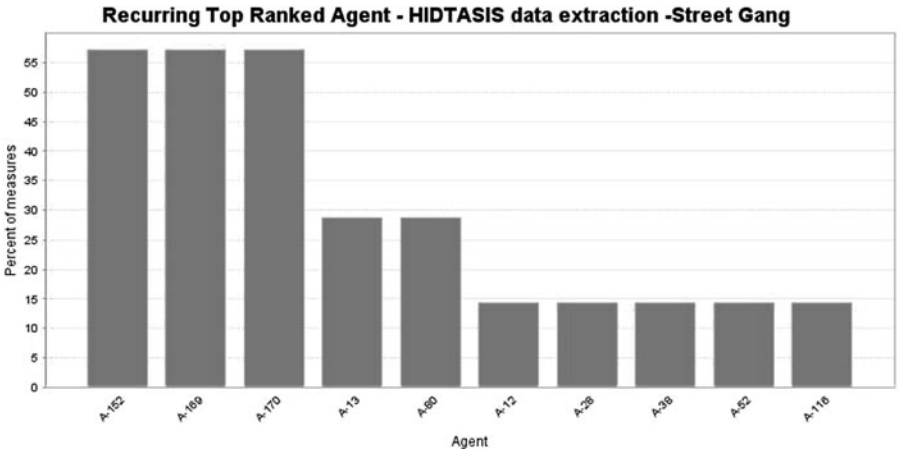


Fig. 5 Top ranked agent chart from ORA Key Entity report

Fig. 6 Group connector rankings table from ORA Key Entity report

Connects Groups (high betweenness and low degree)

The ratio of betweenness to degree centrality; higher scores mean that a node is a potential boundary spanner.

Input network(s): GangOne

Rank	Value	Agent
1	0.110976	A-13
2	0.102592	A-80
3	0.0961693	A-28
4	0.0848062	A-12
5	0.081623	A-68
6	0.0701966	A-55
7	0.0580438	A-58
8	0.0469878	A-48
9	0.0413282	A-29
10	0.0368337	A-97

path to many other nodes, including non-adjacent entities. Leveraging these measures in combination is demonstrated in Figs. 8, 9, and 10. Figure 8 is a chart identifying the 10 nodes which are highest in *Clique Count*. Comparing the nodes in the *Clique Count Ranking* with the nodes in the *Betweenness Centrality* table (Fig. 9), node A-4 is identified as the only node that overlaps both measures.

Figure 10 is a three dimensional view of node A-4's sphere of influence at radius of 1 (direct connections only). Seemingly unremarkable at this network distance, when fully expanded, this node reaches 103 other nodes in the network. This range is shown visually in Fig. 11. This network position and influence make a strong case to promptly investigate or interview node A-4 if it appears that the gang has become involved in an organized action or such action is anticipated by law enforcement.

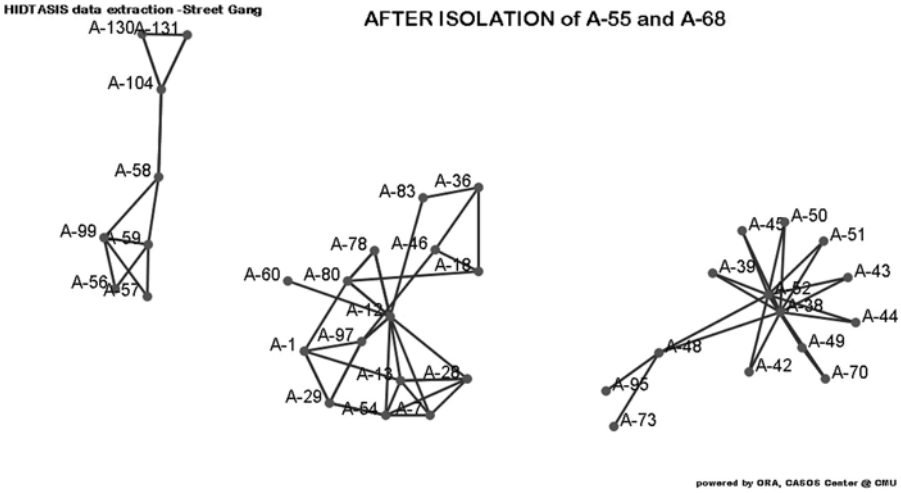


Fig. 7 Gang subgroup after intervention

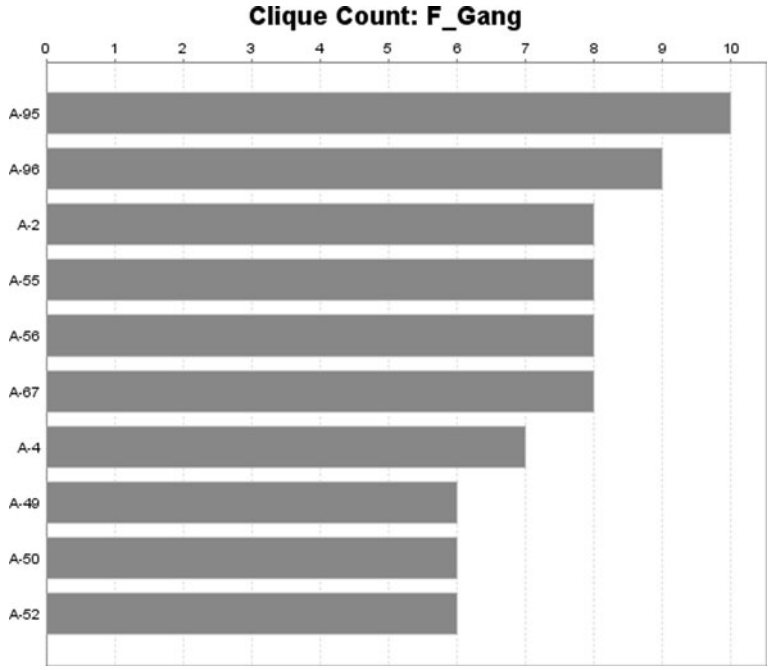


Fig. 8 Clique count chart from ORA "Measure Charts" or "View Charts" option

Figure 12 is an example of applying network analysis to multimodal networks. This table shows that our example network intervention, removal of A-55 and A-68, resulted in an impact on resources. Isolating A-55 disconnected A-95 from the main body of nodes, eliminating or restricting access to resources (R-6 and R-7). This

Fig. 9 Portion of Potential influence ranking table from ORA Key Entity report

Potentially Influential (betweenness centrality)

The Betweenness Centrality of node v in a network is defined as: across all node pairs that have a shortest path containing v, the percentage that pass through v.

Input network(s): F_Gang

Input network size: 215

Input network density: 0.0144534

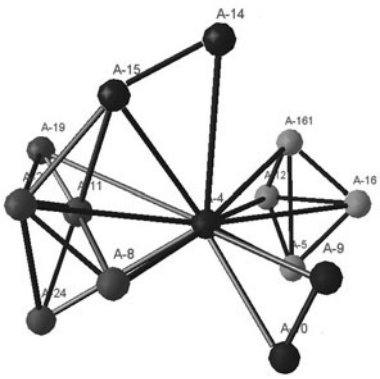
Expected value from a random network of the same size and density: 0.01497

Rank	Value	Unscaled	Agent	Context*
1	0.0282714	1288.67	A-4	0.625821
2	0.0213754	974.333	A-10	0.301369
3	0.0184978	843.167	A-161	0.165981

*Context - Number of standard deviations from the mean of a random network of the same size and density

Unscaled - raw calculated data before normalization

Fig. 10 3D visualizer panel showing node A-4 Sphere of Influence, radius 1



impact could be substantial if R-6 and R-7 are weapons or vehicles, and, in fact, both R-6 and R-7 are vehicles. Isolating A-95 immediately after a conflict between gangs/gang members could prevent these gang members from providing the means to retaliation.

Analysis of networks of any size can be enhanced by running the *Standard Network Analysis* report. This provides useful global information about the network and its structure. These data comprise the big picture for the network and can help in understanding whether our information is grossly incomplete, what type of network it is, and/or how quickly it might respond. Both graphical and tabular data are given for the network in this report. This network was shown to be a low density (0.001065), non-reciprocal (0.121) network with low communication efficiency (path = 4.506). Block clustering applied to the data showed numerous isolates and 1 small cluster. This network would be expected to produce randomized and small group acts as op-

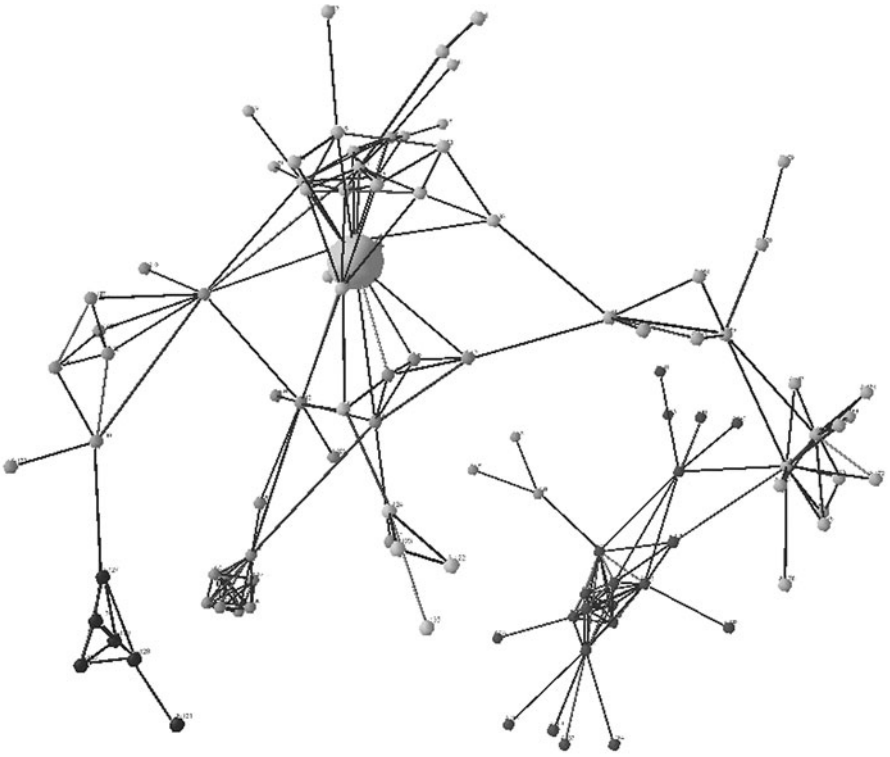


Fig. 11 3D visualizer panel showing node A-4 at full expansion

Fig. 12 Resource Exclusivity table showing node A-95 controlling R-6, R-7 from ORA Management report

Complete Exclusivity - resource (complete exclusivity)

Detects entities that have ties that no other entity has.

Input network(s): agentXresource 3

Rank	Value	Unscaled	Agent	Speciality
1	0.0625	2	A-32	R-16, R-18
2	0.0625	2	A-95	R-6, R-7
3	0.03125	1	A-28	R-28
4	0.03125	1	A-29	R-23
5	0.03125	1	A-31	R-21
6	0.03125	1	A-33	R-3
7	0.03125	1	A-34	R-27
8	0.03125	1	A-35	R-8
9	0.03125	1	A-73	R-4
10	0.03125	1	A-93	R-25

Unscaled: raw calculated data before normalization

posed to well-organized large scale activities. It would have a relatively slow response to aggravating incidents and isolation of a few key individuals could slow responses even further. In the event of civil, gang-related disturbances, police can direct their attention to these few.

7 Network 2

This network is an illustration of the efficiency of analysis introduced by the software. The data import process consolidates entities and records all links between them in a matter of seconds. By applying a transform operation from the main panel, all pendants and isolates can be removed. (Note: this operation takes only seconds on a dataset of this size. The process took less than 20 minutes for more than 2.5 million calls in a recent analysis.) The editor panel in ORA allows nodes to be created, deleted, or merged. The merging function of this panel is very efficient. We were able to reduce the 147,629 calls to 933 nodes in one working day.

This reduction process is iterative and top ranked nodes were evaluated frequently to guide the analysis. Use of these reports focused our efforts in merges and we performed SOI visualizations of the network to avoid obscuring distinctive subgroupings where an agent had multiple phones. These individual phone networks were examined by analysts for any visible differences that might indicate personal social groups. After this review, these phones were merged or deleted. The reviews eliminated agents that were actually public forums such as the national helpdesk of an Internet service provider, which, although high in centrality measures, are not meaningful to the analysis.

With the nodeset reduced to less than 1000, the charts and reports provided by the software were used to select nodes of interest and further discovery. Each of the top ten entities overall were subjected to SOI visualizations, as were the top 10 leaders of strong cliques. This process constituted a re-learning of the network. Many of the network principals remained the same as those identified by decisions of the investigators. The SNA analysis results differed significantly in at least the important discoveries.

The first of these discoveries was a result of the SOI report for node A-1199, selected based on having the highest *Eigenvector Centrality* in the network. Figure 13 shows this report. Since the original telephone data is directional and the import process maintains this property, we found A-1199 had received calls from nearly all the main figures in this investigation but no investigator knew who A-1199 is. Thus, we had a node that had the highest *Eigenvector Centrality* in the network that had not been identified in the previous 2 years of investigation.

Further discoveries resulted from the SOI visualizations; Fig. 14 shows the subnetwork formed by selecting nodes with links to inns and motels (hostelry).

The temporal contexts of these calls show that some of the principals have been using local hostelry over periods of several days. Since these persons are known drug traffickers, we now have additional locations to include as surveillance points in the investigation. The idea to look for a hostelry subnet came from the SOI visualizations predicated by the *KE* top 10 ranked nodes in all SNA measures.

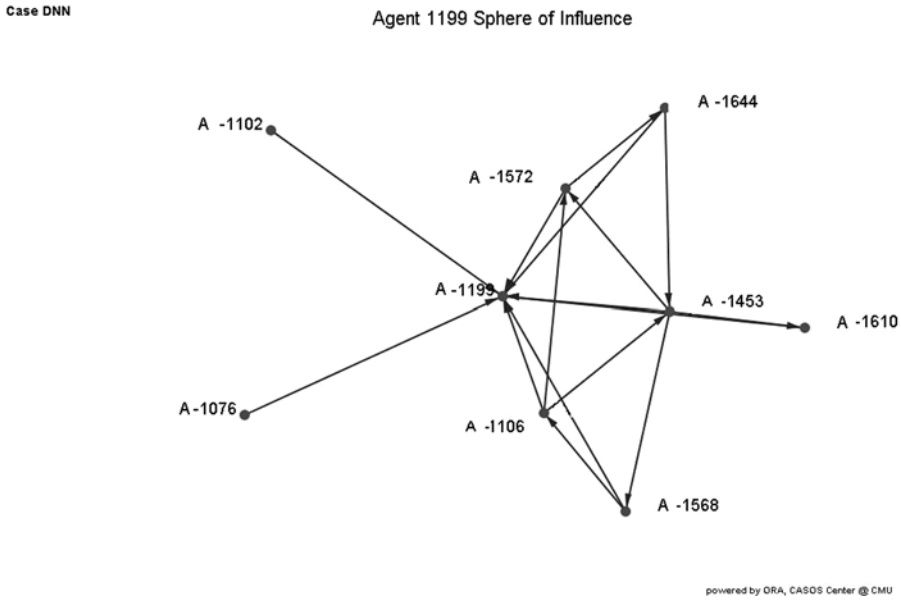


Fig. 13 Sphere of Influence visualization for highest Eigenvector Centrality

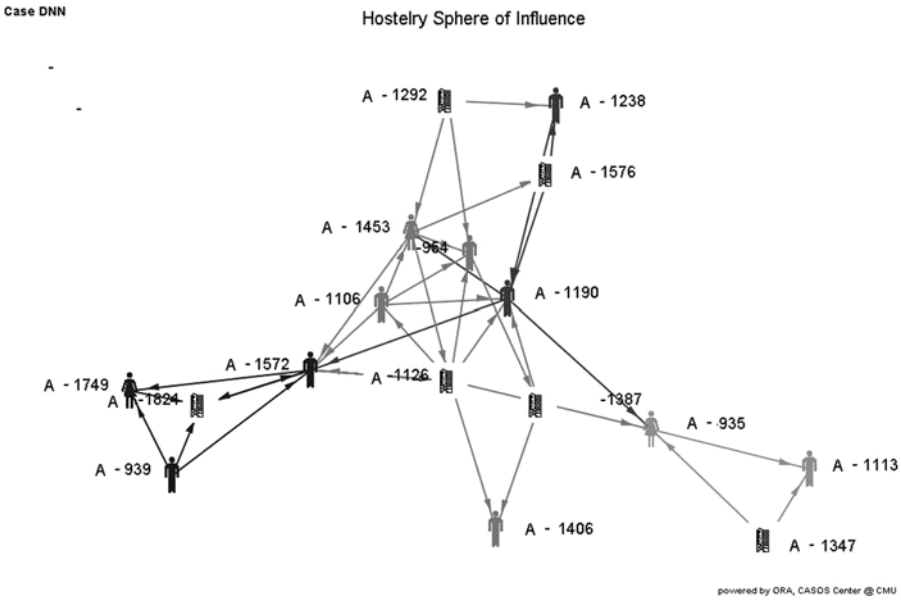


Fig. 14 Nodes linked to inns and motels (hostelry)

Figure 15 illustrates another SOI result. The central node (A-994) was identified as a recurring connection in 3 of the top 4 in a *Recurring Top Ranked Agents* table. In this instance, the visualizations showed a judicial office of the state government as

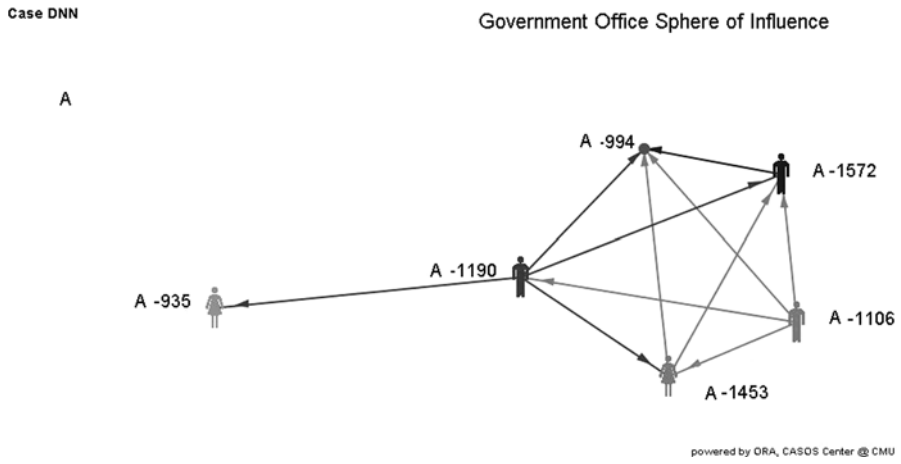


Fig. 15 *Recurring Top Ranked Agent* involved with government agency

a common factor. This may be an important identification of a crossover node, where the illicit network of drug traffickers is aided by a person embedded in the licit world. The ability of SNA to identify this type of node quickly in a nodeset of such size is a critical contribution to our work. Since this discovery we have begun looking for a number of functional groups including financial groups, automobile groups, and service business groups; which increases our understanding of how these traffickers operate. Once again, this information has always been available in tables but rarely acted upon because of data obfuscation.

8 Network 3

Network 3 is an application of criminal network analysis to a Highway Traffic stop case in which more than $100,000 was found concealed in a vehicle (Fig. 16). Although extremely limited in scope, we consider this a building block to be used in aggregating a national dataset for law enforcement. This network was developed from cell phone data files, police affidavits, police car audio transcripts, photographic evidence, and information electronically scanned from papers, receipts, notebooks, business cards, etc. at the scene and on the persons of the suspects. The phonebook and calls information had been extracted from the cell phones and was provided as Microsoft Word and Excel files. We imported this data using the import wizard in ORA. Telephone data was modeled in two ways in the network development: call log items were treated as agents, phone book entries were treated as knowledge. By treating the information in this way agent to agent links are generated where documented calls exist and agent to knowledge links are generated where no documented call exists. This preserves data without claiming unproven links. If data is collected and maintained in this fashion in future cases and is then aggregated, a knowledge to knowledge network will establish indirect ties between agents. The remaining nodesets and graphs were created manually in the editor. The entire process, other than

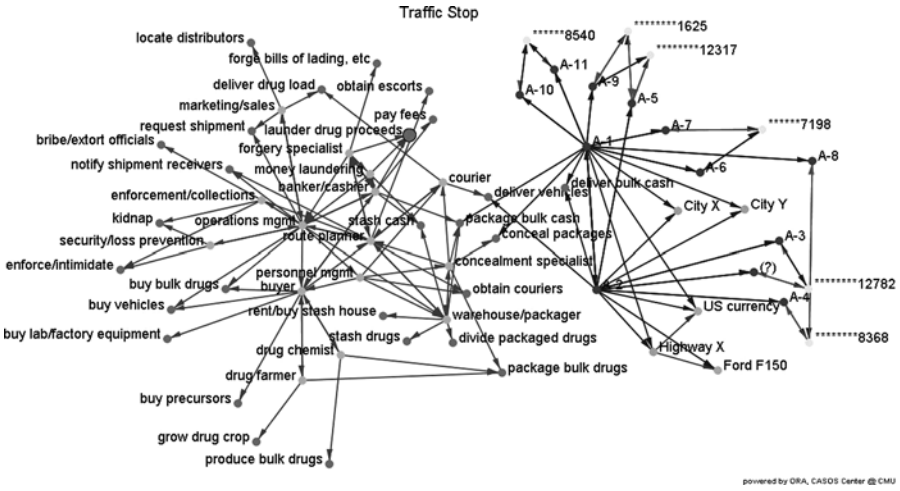


Fig. 16 Multimodal network from highway traffic stop

identifying latitude and longitude data for addresses, was completed in less than 2 hours.

The network is a diagram of everything connected to each of two persons. Placing the physical evidence gathered into a network environment showed numerous connections between the two. Importing the calls and phonebooks identified common telephone numbers they called under different names. Identifying the motels, store locations, tolls paid and the respective date and time information on the receipts established a timeline and route that appears to be involved in the money transportation activity. A limitation of this network is that it does not have sufficient data to make many of the DNA computations useful and the addition of even one additional person could have major effects on computed values. However, the SNA computations do suggest differences between the two that are supported by the physical evidence. Node A-1 is the higher ranked in *Eigenvector Centrality* and node A-2 is ranked higher in *Betweenness Centrality*. The physical evidence in this case showed that node A-1 paid for all of the lodging on the associated route and is the older of the two with a large disparity in age. Node A-2 had more contacts in the telephone evidence and bought the tools and products used in concealing the money. His higher ranking in betweenness is likely caused by the phone information.

The evidence suggests a two person partnership in a criminal enterprise with ties to two other unknown persons who these principals know by different names. The evidence strongly indicates a drug trafficking nexus. We know that the money delivery activity of a drug network is only a small segment of a whole process. For this reason, this network has been tied to a model drug trafficking network to show how the two persons identified likely fit into an overall dynamic network, illustrating the intelligence gaps involved with advancing an investigation. The model used is shown on the left side of Fig. 16. Our traffickers provided vehicles, concealment and delivery of money. There are two shared contacts (node A-4 AKA node A-8) and (node A-5 AKA node A-9) that were identified through their common telephone number

being called by both A-1 and A-2. These would be logical leads to try to discover their role(s) in the money transport activity. If these agents were only points in the money transfer process it doesn't follow that both nodes A-1 and A-2 would have contact with them. Therefore, it seems a better probability that these are brokers in a larger network that has an authority over both agents.

9 Summary

Every criminal network analysis we have presented to law enforcement investigators of their own data has found bridges between groups in the data that startled them. The computational view of data is new to law enforcement analysts though they have used manual visual graphs and computerized tables for a number of years. Mapping software has recently been recognized by many departments as a useful tool in investigation. Most information data sets we see are composed of data contributed by independent observers, obfuscated instance information, disparate data types, and constrained by archaic schema. It is not surprising that these bridges go unrecognized in manual analyses.

The ability to fuse the information streams in these investigations rapidly and accurately has the potential to shorten the duration of drug trafficking cases and aid law enforcement practitioners in recognizing patterns in the actions and resources that traffickers use. We have found that complex cases can be quickly focused by knowing the network level measures provided by SNA, and that the more we explore the relations embedded in the networks, the more utilitarian the explorations become. In addition, the SNA measures take telephone records past the emphasis on call frequency, where large call volume is interpreted as a lead, and replace it with the centrality measures of betweenness, closeness, and eigenvector computations which place importance on network position and the value of connectedness. These expanded measures, of which the networked individual may not even be aware, enrich our ability to understand why the network operates successfully and where precise intervention may produce more return on our efforts. As we understand the person-to-person relationships we can also investigate the means of criminal networks, i.e., the financing, transport supply, and recruiting processes that facilitate the drug traffickers. When we expose this information to the DNA software, measures are computed and rankings developed for these resources as they relate to the actors in the network and to each other. These network properties have only been apparent to those most closely involved in the investigations and may not have been effectively shared in the past. Application of DNA software, such as ORA, for intelligence practice is in its infancy yet we have consistently seen the advantage of clearly identified bridges, ranked clique leader lists, ranked communicators, leveraged resources and exposed links to the licit world.

The ORA program is the only software we have found to date which also provides measures beyond traditional SNA computations. Among these are Cognitive Demand and various Congruence measures. These measures offer opportunity in enhancing criminal network understanding and investigation where data has been collected to support the computation. We expect analysts that use ORA to ask investigators to

expand the information they collect to include knowledge, task and role data so these computations can be used to enhance interventions by identifying new weaknesses in the network.

The concept of "intelligence led policing" and the National Criminal Intelligence Sharing Plan have been increasingly advocated since the September 11 attacks. DNA is an important element in the implementation of this plan by placing a user friendly third generation tool, which can respond dynamically to new information and aggregate case data rapidly, in the hands of law enforcement analysts.

It is rare to examine a dataset using DNA without learning something new about the exposed network, the analysis process, or the meaning of the data and computations. Although the outcome of the mathematical computations is always the same for a particular set of data, the analyst's experience will combine with the information available through visualization and quantitative analysis in new ways. An analyst can easily partition networks, aggregate small networks, or identify common factors in disparate networks by looking at the intersection of those networks. By exercising some of the nearly infinite choices offered for identification and exploration of criminal networks by the ORA software, the analyst will become intimate with the data in ways that a two-dimensional table can never provide. These factors will combine to expose criminals and/or criminal enterprises that are hiding in plain sight.

Appendix: Using charts and reports

The four reports we most often use are the *Key Entity* (*KE*) report, the *Standard Network Analysis* report, the *Communicators* report, and the *Local Patterns* (*LP*) report. In addition, once the individuals are ranked, we have found the *Sphere of Influence* (SOI) report to be of great interest for selected nodes, especially the resource leverage diagrams. These diagrams show a "who has what" view. For example, if your subject is about to flee, it could show vehicles he/she could commandeer or where the subject might try to hide.

The *Key Entity* (*KE*) report is one of the most useful reports for drug trafficking organization analysis. It provides rankings of individuals in the network based on the major SNA and DNA metrics, which show the strongest influences on network behavior based on its structure. This report clearly identifies the difference between simple link analysis and DNA—when the effects of connectivity outweigh the effects of activity. A full *KE* report consists of "Who", "What", "Where", "How" and performance indicators. The "Who" report identifies agents of interest by ranking agents against specific measures such as Cognitive Demand, Eigenvector Centrality, and Betweenness Centrality, in categories that analysts recognize such as *Emergent Leader*, *Leader of Strong Clique*, and *Potentially Influential*, respectively. The "What", "Where", and "How" computations create similar classifications for knowledge, location, and task/resource entities. The performance indicators for a typical network include the sizes of entity classes, the complexity, social density, social fragmentation, and average communication speed of the network. If this report is run before and after an intervention, changes in these values can indicate a successful intervention, or show no change in spite of large changes in other measures.

The *Standard Network Analysis* report provides some of the same measures as the *KE* report but provides additional measure results for some important network facts. The *Exclusivity* and *Specialization* tables show imbalance in network sharing of knowledge, resources, etc. The *Communicators* report provides a table of ranked nodes in three directional measures: *Betweenness*, *In Degree Centrality*, and *Out Degree Centrality*. These measures summarize the network in terms of how information or resources may flow through it. Once the role/roles the individuals play are established, targeting these nodes in the investigation may improve the effectiveness of assigned law enforcement resources.

The *Local Patterns* (*LP*) report provides lists where other reports provide rankings. In particular, the clique analysis list has been helpful, the nodes with hidden links, and the star distribution chart help in selection of *Sphere of Influence* (SOI) targets. In large networks, any process that quickly identifies individuals of interest saves investigative resources. Most often we find interest in the nodes that are different in some manner. For example, the charts provide a graphical display of the range of network measures for quick review; by selecting the betweenness centrality measure in the list of charts, we see a vertical bar graph of that metric. The shape of the curve made by the endpoints of the bar chart shows the variability of the measure. The more exponential that endpoint is in appearance, the more rapidly its maxima decrease. A straight vertical line of endpoints indicates that metric will probably not be helpful in finding unique, influential nodes in the network. In addition, histograms are useful for identification of polarization in the network; if two peaks appear there may be a reason to investigate why one part of the network dominates another. Histograms with equal values in all bins suggest an information gap in the data. Analysis of these tables and charts provide insight into how the network may react to interventions, or show where isolating some resources can influence the network's ability to complete some of its goals. For the examples provided in this paper these reports provided a list of individuals that were targeted for further investigation and proved to be excellent sources of information when planning interventions.

References

Boissevain J (1974) Friends of friends: networks, manipulators and coalitions. Blackwell & Mott, Bristol

Carley K, Reminga J (2004) ORA: organization risk analyzer (Technical Report No. CMU-ISRI-04-106). Carnegie Mellon University, School of Computer Science, Pittsburgh

Carley KM, Diesner J, Reminga J, Tsvetovat M (2007) Toward an interoperable dynamic network analysis toolkit. Decis Support Syst 43(4):1324–1347

Coles N (2001) It's not what you know it's who you know that counts. Analyzing serious crime groups as social networks. Br J Criminol 41(4):580

Diesner J, Carley K (2004) AutoMap 1.2—Extract, analyze, represent, and compare mental models from texts. (No. CMU-ISRI-04-100). Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Pittsburg

Diesner J, Carley K (2005) Revealing social structure from text: meta-matrix text analysis as a novel method for network text analysis. In: Narayanan VK, Armstrong DJ (eds) Causal semantic networking for information systems and technology research: approaches, advances, and illustrations. Idea Group, Harrisburg, pp 81–108

Diesner J, Carley K (2008) Conditional random fields for entity extraction and ontological text coding. Comput Math Organ Theory 14:248–262

Innes M, Fielding N, Cope N (2005) "The appliance of science"? The theory and practice of crime intelligence analysis. Br J Criminol 45(1):39–57

Klerks P (2001) The network paradigm applied to criminal organizations: theoretical, nitpicking, or a relevant doctrine for investigators? Recent developments from the Netherlands. Connections 24(3):53–65

McGloin J (2005) Street gangs and interventions: innovative problem solving with network analysis. Department of Justice, Office of Community Oriented Policing Services, Washington

Ozgul F, Bondy J, Aksoy H (2007) Mining for offender group detection and story of a police operation. Paper presented at the Proceedings of the sixth Australasian conference on data mining and analytics - Vol 70. From http://portal.acm.org/citation.cfm?id=1378245.1378270#

Ressler S (2006) Social network analysis as an approach to combat terrorism: past, present, and future research. Homel Secur Aff 2(2):1–10

Stewart TA (2001) Six degrees of Mohamed Atta. Business 2.0 2(10):63

Williams P (2001) Transnational criminal networks. In: Arquilla J, Ronfeldt D (eds) Networks and netwars: the future of terror, crime, and militancy. Rand, Santa Monica, pp 61–97

Xu J, Chen H (2005) Criminal network analysis and visualization. Commun ACM 48(6):100

Xu JJ, Hsinchun C (2005) CrimeNet explorer: a framework for criminal network knowledge discovery. ACM Trans Inf Syst 23(2):201

Christopher E. Hutchins is the Information Technology Manager for the North Texas High Intensity Drug Trafficking Area (HIDTA) which falls under the direction of the Office of National Drug Control Policy. He received his B.S. in Psychology from Northeastern University. He has been a solution provider to management by introducing efficiency to large tasks and critical analysis of data and programs in the fields of nuclear engineering and construction, computer network systems, and currently specializing in law enforcement. His responsibilities include locating and implementing technological improvements in the areas of detection, investigation, and response to crime, with a particular emphasis on drug trafficking. He works closely with local and national police intelligence managers and analysts in strategic and tactical planning. He is also the developer of the High Intensity Drug Trafficking Area Strategic Information System (HIDTASIS).

Marge Benham-Hutchins is an Assistant Professor at Northeastern University in Boston, MA. She is an associate of Bouvé College of Health Sciences and the School of Nursing. She received her Ph.D. from the University of Arizona with a focus on healthcare informatics and nursing systems and a Master of Science in Nursing (MSN) from the University of Texas (Arlington) with a focus on administration and management information systems. She is also a registered nurse with a clinical background in nephrology and oncology. Her research interests include the use of social network analysis (SNA) to examine the influence of health information technology on communication of patient information between health care providers from multiple professions. Her most recent work incorporated complexity science principles and SNA methodologies to examine communication of patient information between the providers responsible for the transfer (handoff) of a patient from the emergency department to an inpatient hospital unit.