Prototype Geospatial Data Integration Framework for Police Blotter Crime Analysis

Raytheon-UTD Collaboration

1. OBJECTIVES

The overall objective of this proposal is to model and mine geospatial (GS) patterns in multi-jurisdiction and multi-temporal (MJMT) datasets to accurately track, monitor, and predict human activities. Several scientific and technical challenges arise when modeling GS patterns in MJMT datasets due to the spatio-temporal nature and heterogeneity of datasets. These represent major barriers to progress in the Geospatial Information Science (G.I.Sc.) fields such as environmental criminology.

Thus, we propose the following:

To develop a prototype geospatial data integration framework for MJMT data sets to conduct crime analysis based on Police Blotters

Our future work will include pattern analysis, social network analysis, security, uncertainty reasoning

2. CURRENT STATE OF POLICE BLOTTER CRIME ANALYSIS AND LIMITATIONS

Current state of knowledge may be categorized into two main areas: environmental criminology and geospatial data analysis. In environmental criminology, there are several GS theories such as Routine Activity Theory (RAT) and Crime Pattern Theory (CPT). RAT suggests that the location of a crime is related to the criminal's frequently visited areas. CPT extends this theory on a geospatial model shown in Figure 1. This geospatial model consists of nodes (e.g. frequently visited areas such as home, work, entertainment/ recreation), paths (e.g. routes between nodes), and edges (i.e. boundaries of an activity footprint). CPT suggests that crime locations are often close to the edges, i.e. near the criminal's activity boundaries, where the residents may not recognize the criminal. Geospatial data analysis techniques such as geospatial statistics and geospatial data mining have explored geospatial regression, geo-space-time interaction (e.g. Knox test), geospatial clustering for hot spot detection, geospatial outliers, co-location of subsets of crime-types, etc.

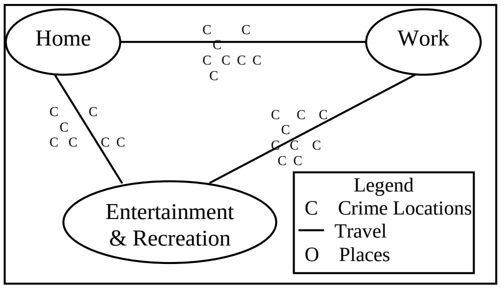
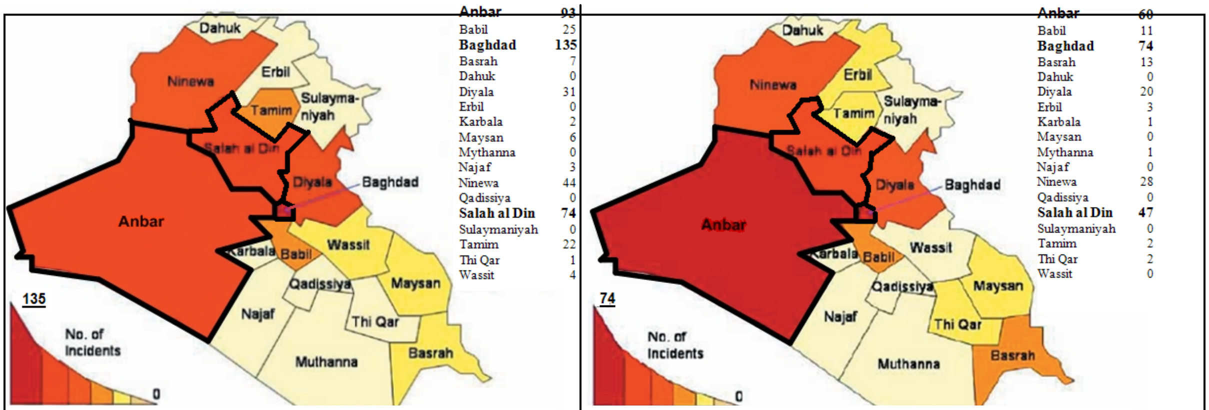


Figure 1: Crime Pattern Theory

The critical barrier of modeling the MJMT data heterogeneities and rich pattern semantics limits environmental criminologists (e.g. many at the Ninth Crime Mapping Research Conference, 3/2007, http://www.ojp.usdoj.gov/nij/maps/pittsburgh2007/index.html) from quickly identifying many GS patterns, which are crucial for timely intervention for crime prevention. There are three significant limitations of current environmental criminology and geospatial data analysis techniques creating this critical barrier. Let us look at these in further detail.



Anbar	**93**
Babil	25
Baghdad	**135**
Basrah	7
Dahuk	0
Diyala	31
Erbil	0
Karbala	2
Maysan	6
Mythanna	0
Najaf	3
Ninewa	44
Qadissiya	0
Salah al Din	**74**
Sulaymaniyah	0
Tamim	22
Thi Qar	1
Wassit	4

Anbar	**60**
Babil	11
Baghdad	**74**
Basrah	13
Dahuk	0
Diyala	20
Erbil	3
Karbala	1
Maysan	0
Mythanna	1
Najaf	0
Ninewa	28
Qadissiya	0
Salah al Din	**47**
Sulaymaniyah	0
Tamim	2
Thi Qar	2
Wassit	0

(a) Jul 19 to Jul 26, 2004 *(b) Jul 26 to Aug 2, 2004*

(Source: http:// www.diligencellc.com) (Best Viewed in Color)

Figure 2: Activity Levels by Jurisdiction (Caution: Use numeric activity count data on right for trend analysis. Color-codes are not directly comparable across Figures 2a and 2b**)**

First, traditional approaches do not explicitly model temporal semantics such as trends or periodic patterns. For example, Figure 2, particularly the numeric activity count data, shows a diminishing trend for the number of insurgent incidents across multiple provinces from July 19-26, 2004 (Figure 2a) and July 26-Aug 2, 2004 (Figure 2b). Notice the highlighted entries in numeric activity count data in Figure 2, e.g. Anbar, where the number of insurgent incidents diminished in a matter of weeks. Timely identification of such GS patterns is crucial for improving public safety. However, it takes enormous amount of time and human effort to identify GS patterns using current tools and techniques, particularly for MJMT datasets.

We are particularly interested in Policy Blotter Crime Analysis. Police Blotter is the daily written record of events (as arrests) in a police station which is released by every police station. These records are available publicly on the web which provides us wealth of information for analyzing the crime patterns across multiple jurisdictions. The Police Blotters are available to public or between police departments are generated from legacy systems and may also be published as web documents. There are major challenges that a police officer would face when he wants to analyze different police blotters to study a pattern (e.g., a spatial-temporal activity pattern) or trail of events. There is no way a police officer can pose a query where query will be handled by considering more than one distributed police blotters on the fly. With the advance of Web 2.0 , there are some mashups of Google Maps with police blotters of some counties. There is not a cohesive tool for the police officer to view the blotters from different counties, interact and visualize the trail of crimes and generate analysis reports. The Blotters can currently searched only by keyword through current tools and does not allow conceptual search, and fails to identify spatial – temporal patterns and connect various dots/pieces. Therefore, we need a tool that will integrate distributed multiple police blotters, extract semantic information from a police blotter and provide seamless framework for queries with multiple granularities.

3. PROPOSED APPROACH

To address the limitations discussed above we will transfer the research we are conducting for Raytheon as well as augment this research by accomplishing the following in developing fully fledged prototype systems. We will use Policy Blotter as our application.

Motivating Scenario:

Police Blotters are available from legacy based systems which causes the data integration problems. The Blotters may come in different data formats like HTML, PDF. Semantic Web Service Interface provides us with the capability to integrate these varied data formats and semi automate the process of integrating different data sources for a unified view. Also the information regarding the crime reported through police blotters are in format not cohesive for machine to interpret for drawing inferences and assertions which are necessary for a scenario mentioned below.

Here we consider the real event that occurred very recently "The Shootings at Virginia Tech" which has raised again the consideration of robust emergency response tools in the hands of the Police to take actions and handle the emergencies. Police blotters of a university crime are available with different University Police departments, and also the blotters from counties of major City like Dallas [data set 1] needs to have an efficient way to integrate the information to analyze the patterns and produce a trail of similar events that help to catch the suspect faster/quickly.

Geo-Spatio-temporal Data Integration

Many environmental criminology techniques assume that data are locally maintained and the dataset is homogeneous as well as certain. This assumption is not realistic as GS data is often managed by different jurisdictions and therefore, the analyst may have to spend unusually large amount of time to link related events across different jurisdictions (e.g., the sniper shootings across Washington DC, Virginia and Maryland in October 2002).

A major challenge that needs to be addressed when integrating heterogeneous crime data sources is semantic heterogeneity. Semantic heterogeneity occurs when there is a definition-mismatch across MJMT datasets (e.g., robbery is a kind of crime). Naming heterogeneity may also exist (e.g., theft and burglary are two different terms but convey the same semantic meaning or concept). Current data-integration approaches (e.g., GML, wrappers, GS ontologies) do not adequately model Police Blotter concepts and heterogeneities. Lack of integrated geospatial-based police blotter activity datasets will make it tedious to perform analysis for MJMT activity patterns. Thus, we propose the development of GS ontologies as a first step toward meeting the challenges in integrating heterogeneous data sources. These GS ontologies will be integrated with other ontologies that provide, for example, definitions of various environmental criminology terms.

We propose dynamic integration of MJMT data through Semantic Web services (SWS). SWS framework allows intelligent data retrieval by annotating the WSDL (Web service description language) profile of agencies' data dissemination point. We have developed a prototype for dynamic geospatial data integration task that is capable of performing responding to complex client queries. The prototype will be enhanced to handle non-geospatial data sources such as criminal data (or money transaction audit). We also plan to augment the prototype by building pluggable automated tools that will relieve field agents of having to perform complex queries. The primary interface to the field agents will contain pre-completed queries, which can be run against each individual. However, managers to the field agents who have better domain knowledge can still utilize a sophisticated query interface.

Our solution makes the assumption that the data from the agencies is encoded in one of the commonly used formats (e.g., PDF, XML, HTML, Relational, RDF). The integration will be a two-step process. In the first step, SWS component retrieves information from each data source with a very high accuracy level. Since these information are still disparate (e.g., an XML file containing criminal history cannot be immediately combined with a Oracle database containing denied persons list), another step has to seamlessly combine them to construct a decision. This is the data mapping part of the integration process. Two methods of data mapping will be used in the second step: ontology method and adaptor method. Our ultimate goal is to map all the retrieved data in a single, unified format. Ontology is formal description of domain concepts and their mutual relationships. Using ontology, we can map concepts from disparate domains into the unified format. If there is no suitable ontology for a domain, an adaptor (also called a converter or wrapper) is built into the system to convert the source data into the unified format. Adaptors will be used to translate legacy database into our target format.

4. Datasets Available:

1. Police Blotters for Dallas County available online http://www.dallasnews.com/sharedcontent/dws/news/city/collin/blotter/vitindex.html
2. Semantic Access Ontology, GRDF Ontology, Geospatial Services Ontology.
3. ClearForest Semantic Web Services for Text Mining and Analysis: http://sws.clearforest.com/ws/sws.asmx?wsdl

5. TASKS AND DELIVERABLES

Task 1: Semantic Search Browser for Police Blotters:
To Provide a Semantic Level Browser that integrates the blotters from various counties or by geographic regions which provides an interface to the police office to query to get information with different input criteria like (a) by Crime Types (e.g., rape case) (b) By Time Period (e.g., in the first week of April 2007) (c) By Suspect Personal information (e.g., crime activity of Mr. X) (d) by geographical region using Zip Codes, City (e.g., list all sex offenders in City of Dallas). We have already developed a semantic framework DAGIS which can handle queries of this nature. The Blotters will be exposed through Web Services for general input criteria and Semantic Web Services of these exposed web services will provide the capability to do more conceptual searches and dynamically compose services on the fly to handle various complex queries. Therefore, integration problem of multiple police blotters will be addressed. In addition, this solves semantic heterogeneities across blotters and provides an automated discovery of knowledge.

Task 2: Tools for Generating Crime Analysis Concepts from Blotters
Information available in the blotters would be mined by developing using Data Mining tools which would be used to generate the concepts that would be mapped to build an Ontology for Crime Analysis across multi-jurisdictions. These tools will also be exposed as Semantic Web services and can be integrated with the Semantic Search browser developed in Task 1. In DAGIS we have developed OWL-S based semantic web service for the ClearForest Text Mining

Semantic Web Services[3]. In future, we would like to develop techniques that will generate semantic representation of concepts and their relationship given a police blotter report.

Task 3: Map based Visualizing Tools and Semantic Dashboard

Current blotters are usually text based with some overlay (e.g., Map mash up) but there is no visualizing tool that provides a trail of crimes occurred according to the search criteria (i.e, fail to connect dots). However, if we can connect these dots using some visualization tool, the Officer will be able to make decisions quickly and efficiently during emergencies. We would like to develop a Map based Visualizing display for overlaying the results of the blotter search and the analysis. This output of display will be a semantic dashboard like wiki based pages.

Task 4. System Demonstration:

We will design and demonstrate a full fledged prototype system that will utilize real world datasets from multiple jurisdictions. (will work with Raytheon on this).

5. References on Crime Analysis:

1) P. L. Brantingham et al., Environmental Criminology, Waveland Press, 1990.
2) N. Levine, "CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations," 3.1 ed. Houston, TX: Ned Levine & Associates, 2007.
3) O. Schabenberger et al., Statistical Methods for Spatial Data Analysis, Chapman & Hall, 2005.
4) F. Wang, Geographic Information Systems and Crime Analysis, Idea Group Publishing, 2005.

APPENDIX A Datasets and Future Work:

Currently we have access to two sets ([1], [2]) of denied persons list from US Department of Commerce and US Department of Treasury, respectively. In a live environment, field agents will have the credentials to access lot more datasets than those disseminated for public or research. There is no need for obtaining the inaccessible datasets as long as the data descriptions (i.e. metadata) are available. The descriptions will be used to create ontology or adaptors. We will work with Raytheon to obtain data sets relevant to our application. This proposal has focused on crime analysis. We will be flexible to handle applications such as Border Patrol and Homeland security.

In the future, we will also support the following features:

- Decision Support Tools/ Crime Pattern activity tools: Decision support involves recognizing patterns, tactics, management planning etc. We will provide automated tools to allow intelligence analysts to review and scrutinize available data. For example, US Customs and Border Protection (CBP) managers can look at the net traffic flow across the border ports and make a decision to divert more resources to overloaded checkpoints and entries.
- Social Network Tool: A sophisticated intelligence analysis requires not only explicit data but association knowledge that can help identify malicious intents in advance.
- Visualization Tool: Our system will allow analysts to trace a person of interest over a geographic area in a map or through a social network graph.
- Security: We will incorporate policy management tools into our prototype.
- Data Uncertainty Management for Multi-Modal Transportation Systems: We will investigate models to represent diverse uncertainty representations across GS network datasets. These techniques will subsequently be integrated into our data integration framework and tools.

[1] http://www.bis.doc.gov/ComplianceAndEnforcement/ListsToCheck.htm
[2] http://www.bis.doc.gov/dpl/Default.shtm

APPENDIX B Homeland security application

US border ports are critical for a stable economy and cross-border trade. Annually US Customs and Border Protection (CBP) process over 25-million (70,000 daily) cargo entries through the 149 ports in the United States. Out of them, 6.7 million entries come in trucks via 14 border states. The current mode of security in cross-border highway trade is implemented by manual inspection of the documents and cargo shipments by the field agents. The explosive growth in the influx of truck-transported foreign goods has become a major safety concern because of the outdated security systems. Hand-checking such a large amount of truck shipments creates security vulnerabilities and vehicle backlog.

Department of Homeland Security has initiated the ACE (Automated Commercial Environment) electronic manifest system in an effort to mitigate the problem. ACE e-manifest would allow carriers to send a manifest of the vehicle and products on-board to the CBP in advance. While this will expedite trade flow across the ports, security concerns remain. Individuals boarded on the vehicles are cleared by running their identification through only a small set of datasets. Despite the availability of many other potentially pertinent databases maintained by state and federal agencies, lack of interoperability among the data sources is a severe obstacle. The non-interoperability issues get in the way of data integration and subsequent decision support mechanisms.

Motivating Scenario: A carrier sends the e-manifest stating the vehicle personnel and company information through the broker. One of the persons in the list is a human smuggler and has used multiple aliases in the past. The ACE system will be of little help in determining the threat and currently there is no automated security mechanism in place to provide warning signs to the field agents. Assuming the agent somehow manages to discover probable links, there is no secondary system to provide concrete actionable intelligence for the field managers.