Intelligence Database Creation & Analysis: Network-based Text Analysis versus Human Cognition

John M Graham, LTC, PhD
John.Graham@us.army.mil
Center for Network Science
Engineering Psychology Laboratory
United States Military Academy

Kathleen M Carley, PhD
Kathleen.Carley@cs.cmu.edu
Center for Analysis of Social and Organizational Systems
Carnegie Mellon University

Drew Cukor, LTC, MS
United States Marine Corps

Abstract

The 9/11 Commission Report and the National Intelligence Reform Act both state that the development of terrorist network database collection processes is an immediate and pressing requirement. This paper is a study and comparison of two complementary approaches to developing a terror network dataset: Automap [15] a Network Text Analysis (NTA) tool; and Intelligence Analyst coding, a human process. NTA tools are an emerging branch of software that supports the analysis of quantitative characteristics of large-scale textual data [24] as well as the extraction of meaning from texts[6]. Intelligence Analyst coding is the traditional method that requires a human to read and cognitively process each raw field report. In this study, both approaches were applied to the same one hundred eighty-three open source texts on the Al Qaeda organization. Each approach's process, dataset product, and analytics are compared qualitatively and quantitatively. In terms of process, the Automap-assisted system required less manpower and time resources. In terms of dataset product, both approaches identified unique nodes and relationships that the other missed. Lastly, the differences in the datasets significantly impacted threat analytics and potential course of action selection. These results suggest an integrated human-centered automation support approach to intelligence dataset development.

1. Introduction

While intelligence agencies use multiple methods to conduct threat assessment, all methods' validity are, ultimately, dependant on the thoroughness of the threat datasets [25]. In counter-terrorism, a normative threat dataset consists of a collection of known probable and possible people with terrorism connections, their attributes, and their relationships [8]. Improper assembly of raw data into datasets has been cited as an issue by the Department of Justice and the FBI [13,21]. In fact, the 911 Commission Report repeatedly commented that when good analysis did occur in the US intelligence agencies, it was on incomplete or poorly constructed datasets (9/11 Commission Report, 2004).

As a result, dataset generation and sharing is recognized as a key area for improvement in the U.S. intelligence apparatus [22]. Multiple processes exist or are being developed to turn raw data into analyzable datasets. These approaches range from the pure human-cognitive processes which are done by a trained analyst to fully-automated processes that require no human in the loop (figure 1).

Three examples are the target folder methodology, software-guided methodology, and Automap-assisted process. In military intelligence, the target folder methodology is used to organize data by a human analyst [1]. In law enforcement intelligence, link analysis software is often used to structure and assist the human coding [18], but this process still requires the analyst to read and assess all raw field reports. And, recently, a few software systems, such as Automap, have been developed to partially automate the human coding process [14] which allows the human to devote more of their time to analysis of the dataset. To our knowledge there are no working examples of fully-automated database building that do not require a human in the loop.

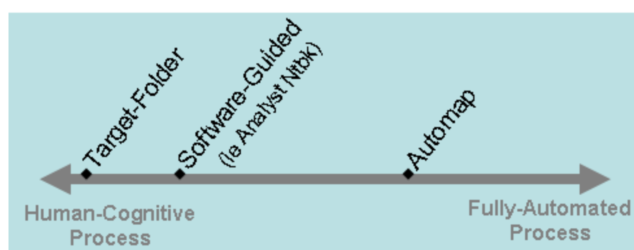

Figure 1. Spectrum of Dataset development approaches. The Target-Folder process is a predominantly human endeavor, while Automap is the closest to an automated endeavor.

Automating the target folder approach has, up to now, not been well received by the intelligence community. Intelligence analysts are used to working with datasets where each data point is referenced by its source and the intelligence analyst that entered the data point. As a result, each data entity, entity attribute, and entity relationship has a pedigree based upon the reputation of the intelligence analyst that input the information. Automation, without a human in the loop, does not have the requisite pedigree.

Unfortunately, pure human-cognitive approach suffers from problems as well. The process of schema mapping, information extraction, duplicate elimination and standardization is tedious and high in cognitive demand [22]. As a result the purely human-cognitive process is subject to errors of omission and errors of commission associated with high cognitive demand work [17, 28].

The sheer amount of field reports is also making the pure human-cognitive process untenable. Faced with an increased amount of raw reports from field agents, the human intelligence analyst can only add the most important reports to the databases. As a result, many potentially important field reports are never coded into databases.

Unhappy with this state of affairs, intelligence science and technology departments recognize that a new system is required to capture all available information in a form usable for threat assessment. One option is to integrate the human-cognitive processes with the partially-automated processes.

This study is a comparison of two complementary approaches to developing a terror network database given the same raw sources. The two approaches are the Intelligence Analyst (Human) and Automap-assisted (Software). The Intelligence Analysis approach is based upon the military target folder intelligence methodology and is an intensely human cognitive process.

The Automap-assisted approach was developed at the CASOS Laboratory and consists of a partially automated process controlled by a human. The Automap software is a network text analysis tool that extracts, analyzes, represents, and compares mental models from texts. Network text analysis is a specific text analysis method that encodes links between words in a text and builds a network of the linked words [14]. Similar text analysis tools include CETA [23] and KEDS [20].

The traditional intelligence analyst approach and the Automap-assisted approach are complementary. As we will demonstrate, the Human Intelligence Analyst process suffers from limitations compensated for in the Automap-assisted approach and the Automap-assisted process suffers from limitations compensated for in the human-cognitive process. As a result, the combination of both outputs will provide a more complete data set that can compensate for the biases of each.

2. Method

Context and Participants

This experiment required multiple, parallel processes (Figure 2). Both conditions, however, used the same raw data sources in text form. Raw data was first gathered, the two dataset development processes were applied, the respective datasets were formalized, and the results underwent Dynamic Network and Threat analysis. Each phase was compared qualitatively and quantitatively.

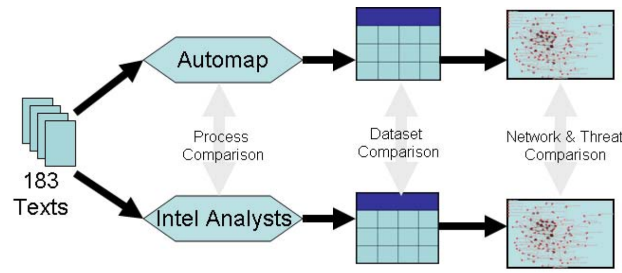

Figure 2. Experimental Design: The same 183 texts were processed through two approaches. The process, datasets, and threat assessments are compared.

The participants in this experiment were novices. For the unassisted analyst condition, three student interns from the Center for Strategic Studies received two days of training on unassisted data extraction from text documents before beginning their tasks. For the Automap-assisted condition, one intern, with intelligence analyst experience was assigned to code and analyze the same data set.

Text Database: The same set of one hundred and eighty three Al Qaeda texts were collected and used in both conditions of this study. Initial set of texts searched from Lexus-Nexus using the twenty 911 attack participants [25]. Beginning with the 911 hijackers, a snowball technique was applied to successively add texts about key AQ members at the time of the 2001 attacks. As a result, the membership expanded to include, 911 hijackers, the Hamburg cell, the Afghanistan & Sudan training cells, Jemaah Islamiya and various individual members without subgroup affiliation.

Intelligence Analyst Approach: Three interns at the Center for Strategic Studies were trained and assigned to conduct the Intelligence Analyst approach to dataset development. The three interns applied military target folder methodology and post-hoc knowledge to assemble inference based agent-agent matrix as well as a set of fifteen attributes for each agent. If a relationship between two agents was inferred, the intern assigned a value of one to the appropriate cell in the adjacency matrix. This process took approximately three weeks with the assistance of significant post-hoc knowledge.

Automap-assisted Approach: Two iterations of Automap meta-matrix generation were used in this experiment: Run1 & Run2. A single intern conducted both Automap-assisted runs. During Run1, the single analyst, using Automap 2.0.10 applied a scan window size of two on the one hundred eighty-three texts. A baseline thesaurus was provided by the CASOS Laboratory from a previous Al Qaeda project. Based on lessons learned during Run1, the analyst changed the Automap scan window size to four during Run2. Prior to Run2, the analyst updated the thesaurus as described in the Automap user manual.

Data Collection & Measures

The two experimental conditions were compared based on the three categories of text to data process, data product, and analysis product. All categories are evaluated qualitatively and quantitatively.

Text-to-Data Process Comparison: Both approaches were compared with respect to the user experience in translating the raw data into inferred relationship matrices. The goal is to understand, quantitatively, the fundamental differences in resources required to achieve a usable database. Qualitatively, we gathered feedback from interviews to characterize the user experience.

Data Product Comparison: The outcome from each data process is a data product. As there is no gold standard or right answer, the data products were compared against one another. Measures included number of same agents in each dataset, number of unique agents in each dataset, and a qualitative analysis of the pros and cons of each dataset.

Threat Analytic Product Comparison: Datasets are collected for the purpose of analysis. We expected that, if there are fundamental differences between the datasets, these differences will result in fundamental differences in the analysis product. Because there is extensive post hoc information about the 9/11 hijackers, we chose to use automated analytics to remove human bias. We applied ORA (Organizational Risk Analysis) software as the standard analytics to the datasets resulting from each condition. ORA is a risk assessment tool for locating individuals or groups that are potential risks given social, knowledge and task network information [21].

3. Results

Text-to-Data Process Comparison:

Quantitative: The Automap-assisted approach required observably less time and resources to complete than the Intelligence Analyst approach. As per Figure 3, the Intelligence Analyst approach required three people approximately three weeks (180 man-hours) to complete. However, Automap Run 1 cost only 19 man-hours. Further, with experience, the same task required only 9 man-hours.

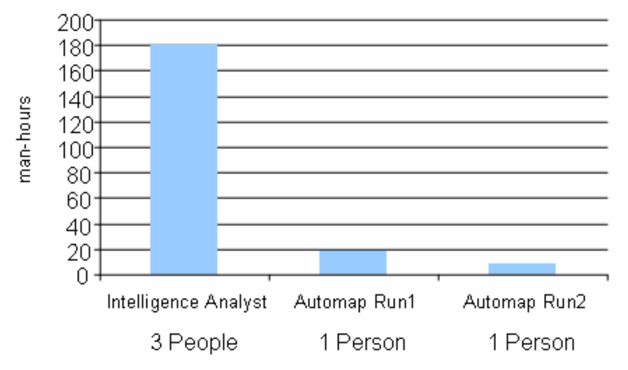

Figure 3. Resource Requirements by Dataset Approach

Automap Run1 was the most time consuming of the two Automap-assisted Runs (Table 1). Almost half of the time (9 hours) was consumed in training. First the user self familiarized with the tool using the online help and a copy of the software. Then the user received abbreviated training (4 hours) on the tool at Carnegie Mellon University. This training is considered abbreviated as the CASOS Summer Institute schedules a total of 12 hours of classroom training on the Automap toolset. Just over half of the novice users time was spent applying Automap to the text dataset.

Run2 showed a significant improvement in time over Run1. Based on the user's experiences, three hours of telephone consulting lead to a more streamlined and accurate data preparation process. As a result, the data preparation and 'crunching' time took 40% less time than Run1.

	Tool Familiarity	Formal Instruction	Application	Total
Run1	5 hours	4 hours	10 hours	19 hours
Run2		3 hours	6 hours	9 hours
Combined	5 hours	7 hours	16 hours	28 hours

Table 1. Automap-assisted Run1 & Run2 Time Break-down

Qualitative: We asked the respective interns to comment on the process of converting text-to-data. The unassisted interns expressed the following concerns about their process: 1) They often lost track of the coding categories and had to return to the definitions, 2) Changes to the coding categories resulted in a restart through the data, 3) they were not convinced that all interns were consistently using the same criteria throughout the process.

The Automap-assisted intern commented the following about his process: 1) the analyst still felt it was a requirement to read the texts to get all of the implicit data contained in the texts, 2) the analyst had no qualms repeating the entire process if the coding or target information changed.

Data Product Comparison:

Terrorist Agents

Intelligence Analyst & Automap Run1: The analysts generated a list of 113 agents from the 183 texts while the Automap-assisted solution produced a list of 85 agents (table 2). 55 of the agents were co-identified by both the analyst and the Automap-assisted methodologies. Therfore, 65% of the analyst list matched the Automap list and 49% of the Automap list matched the analyst lists respectively.

Each experimental condition developed a set of agents unique to the specific approach. In the case of the analysts, 57 of the terrorists listed were unique to their list. A total of 29 agents were unique to the Automap-assisted methodology. Of the 29 unique agents, nine consisted of world leaders or news reporters often mentioned in the context of terrorism and counter-terrorism but were not actual terrorists of interest. Lastly, 4 of the people identified in the Automap thesaurus were repeated agents using an alias.

Intelligence Analyst & Automap Run2: Automap Run2 is compared to the same Intelligence Analyst results. Automap Run2 identified a list of 168 total agents. Sixty-six of the agents were co-identified by both the analyst and the Automap Run2. In this case, 45% of the analyst list matched the Automap list but 67% of the Automap list matched the analysts list.

As in Run1, both lists contained a set of agents unique to each approach. If an agent is unique to one approach, it was missed by the other approach. The unassisted intelligence analysts had thirty seven agents unique to their list. The Automap-assisted Run2 produced a total of seventy-six unique terrorist agents. Of the seventy-six unique agents, thirty-three were the names of world leaders or newsmakers that were not terrorists of interest. Seven of the agents in the Automap Data were found to be alias for other names in the database.

	Run 1	Run 2
Analyst & Automap Agents	55	76
Analyst Unique Agents	57	37
Automap Unique Agents	20	43
Automap US/ World Leaders	9	33
Thesaurus Repeats	4	7

Table 2. Agents Identified from Intelligence Analyst & Automap-assisted runs.

Terrorist Meta-matrix

Intelligence Analysts are interested in more than agents as the only nodes that make up a terrorist organization. In addition to agent nodes, the analysts use other node types and attributes of the agents such as knowledge, resources, tasks, organizations, locations, actions, and roles. These nodes are critical to understanding the context that the agents operate within. In the terms of dynamic network analysis, these additional nodes are considered part of the organizational meta-matrix.

Table 3 is a comparison of the meta-matrix node quantities identified using each analytical method. Clearly, all but one organizational measure, the Automap-assisted method was superior. In fact, the unassisted Analysts chose not to not even collect on many of the categories of meta-matrix data (indicated by n/c). Based upon the comparison, the meta-matrix generated by Automap contains much greater categorical detail that than produced in the unassisted analyst developed data.

META-MATRIX QUANTITIES			
Nodes	Analys t	Automap Run 1	Automap Run 2
Number of AGENT	113	85	168
Number of KNOWLEDGE	9	11	14
Number of RESOURCE	n/c	5	7
Number of TASK	6	0	1
Number of ORGANIZATION	n/c	28	61
Number of LOCATION	n/c	26	55
Number of ACTION	n/c	4	6
Number of ROLE	n/c	1	3

Table 3. Comparison of Meta-matrix Quantities across Analyst and the Automap-assisted Run1 & Run2.

Terrorist Network

To study any network, we must have edges in addition to the nodes (agents, knowledge, resources, etc).

Edges determine the relationships between nodes and provide the pathways for moving information, knowledge and resources. Analysts devote much of their time determining and unpacking the relationships in a terrorist organization. Edges, combined with the nodes, make up the network.

In the target folder methodology an edge, or relationship is established as the analyst sifts through the massive corpus of documents to identify entities (people, places, equipment, etc). This is no simple task, and when left to the unaided analyst, all of the human cognitive limitations begin to work against a reliable outcome (ie boredom, premature closure, inattentiveness, primacy-recency bias, confirmation bias, etc). Further, as the analyst labors through the corpus, their criteria for determining relationships is constantly shifting as they gain intuitions about the culture, individual differences, and meta-analysis of the data set linguistics.

In most applications of target folder methodology, the analyst, familiar with the data set, will establish subjective threshold criteria for classifying an edge either as a relationship or interaction. In this experiment, the novice analysts were instructed to count interactivity greater than one as relationships. Interactivity, one or zero, was rejected as a relationship and not included.

Automap establishes edges based on proximity in a text. In the case of these data sets, edges were established based on a window size parameter of two in the case of Run 1 and four in the case of Run2. Window size is the method Automap uses to establish a relationship between two nodes. It defines how distant concepts can be and still have a relationship. Only nodes within the established window size can be connected by an edge.

As a result, the target folder methodology takes advantage of content of the data, semantic cues that an expert analyst, using informed intuition can then infer relationship over interaction. The Automap approach, on the other hand is a very precise, consistent edge detection methodology. It takes advantage of the relationship evidence of high repetition in a large data corpus.

Full networks can be compared qualitatively and quantitatively. Qualitatively, we are able to visually identify isolates (nodes not connected to the network), pendants (nodes with a single connection to the network), and make general assessments of the density (number of edges divided by the possible number of edges between agent nodes). Quantitatively, we are able to assess the social network density (density of the agentxagent network), overall complexity (number of edges between all meta-matrix nodes), Component Count (Number of undirected components in the organization), and Organizational Network Density (Density of the Organization x Organization network).

Analyst & Automap Run1 Network Comparison: The graphs in figures 4 & 5 were generated in UCINET. The left graph is the agent x agent matrix generated using the Automap data. The right graph is the analyst inference agent_agent matrix.

In figure 4, the unassisted analyst generated matrix has significantly more agents (113) than the Automap-assisted solution (85). Further, there is one third as many isolates in the Analyst generated data. Lastly, visually, it appears that the agent_agent density is significantly higher than the Automap generated datasets.

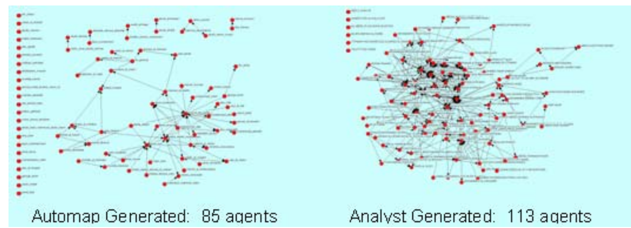

Figure 4 a & b: Run 1 Automap-assisted Data vs Analyst Generated Data

In figure 5, the left graph is the agent_agent matrix generated using the Automap Run2 data. The right graph is the same analyst inference agent_agent matrix compared to both Automap-assisted runs. In this case, the Automap generated solution has more agents (168) than the Intelligence Analyst generated solution (113). Also, very different than the Run1 results, there are fewer isolates in the Automap-assisted dataset than the unassisted analyst generated dataset. Visually, the Automap generated network appears to have fewer central members and more pendants than the Analyst generated agent adjacency network.

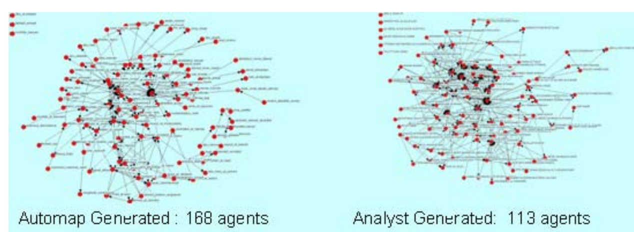

Figure 5 a & b: Automap 2 Generated Data vs Analyst Generated Data

ORA (Organizational Risk Analyzer-CASOS) was used to study the high level statistical qualities of the meta-matrices produced using the two methods (Table 5). Within the ORA options, the Risk Report and Intelligence Report functions generated the values in the table.

In nearly every measure and across all runs, the unassisted analysts produced a richer data set. Essentially, the analysts have more edges in their data relative to the number of nodes than either Automap-assisted solution. In fact, the overall complexity and network density are more than double that of the Automap generated dataset.

Analysis Product Comparison:

ORGANIZATIONAL CHARACTERISTICS			
Characteristic	Analyst	Automap Run1	Automap Run2
Social Network Density	0.0394	0.0126	0.0099
Overall Complexity	0.0584	0.0157	0.0217
Organizational Network Density	N/A	0.0132	0.0164

Table 4. Organizational Characteristics Comparison

The standard ORA Intelligence Report was used to generate an analysis of the key leadership in each organization. A range of analytics that depend upon different node and edge characteristics were chosen for comparison.

The following tables are a comparison of the top five 'leaders' as indicated by their centrality and power measures across all experimental conditions. Centrality measures indicate agents with influence or information within the agent x agent network. Power measures indicate agents with access to assets (knowledge, resources, etc) that necessary for task completion. betweeness, Eigenvector centrality, and Degree centrality indicate different types of leadership or power in the organizational network.

Centrality

Betweeness Centrality: Only considers the agent matrix. This is the person through which information will most likely pass to get from one member of the organization to another. Ramzi's unique place as number 4 in the Automap-assisted data is telling of how much Ramzi wanted to join the 911 hijackers (Table 5). He made multiple contacts in his multiple unsuccessful attempts to enter the United States. As a result, he knows a great deal (and remains in a secret location under US guard). Saud al-Rashid was the unique high betweeness centrality member in the Intelligence Analyst dataset. His multiple roles as a financier and former Afghanistan training camp member provided many relationships that were apparent to the three analysts reviewing the texts. It is interesting to note that the Automap-assisted approach identified Ramzi, an operational terrorist, while the intelligence analyst approach identified Saud al-Rashid, a terrorism financier.

	Centrality-Betweenness		
Rank	Automap Run1	Automap Run2	AnalystOrg
1	UBL	UBL	UBL
2	KSM	KSM	Atef
3	abdul_aziz	abdul_aziz	KSM
4	Ramzi	Ramzi	Saud al Rashid
5	hani_hanjour	Atta	Atta

Table 5. ORA-based Top Five Leaders (Betweenness Centrality) in the Automap-assisted and Analyst Datasets.

Eigenvector Centrality: Only considers the agent matrix. These are the leaders among leaders. Interestingly, the Automap-assisted data includes President Bush, which demonstrates that the measure is effective…for while the UBL and President Bush are adversaries, they are both leaders among leaders. In total only two of the top five matched for the Automap-assisted and the Intelligence Analyst approaches. The leaders identified in the Automap-assisted data set were few while the difference in Eigenvector Centrality between the Intelligence Analyst top ten was insignificant.

	Centrality-Eigenvector		
Rank	Automap Run1	Automap Run2	AnalystOrg
1	UBL	UBL	UBL
2	tawfiq_attash	Zawahiri	Atta
3	abdul_aziz	Atta	Khalid al Mihdhar
4	george_bush	KSM	Ahmad al Haznawi
5	KSM	abdul_aziz	AHMED AL-NAMI

Table 6. ORA-based Top Five Leaders (Eigenvector Centrality) in the Automap-assisted and Analyst Datasets.

Degree Centrality: Only considers the agent matrix. These are the individuals with the most agent to agent communications coming in and going out. If you wanted to spread a rumor or misinformation, this is where you would start. These agents are often the easiest to target… they talk too much! Both data set approaches agreed that KSM, UBL, and Atta were highly connected agents.

	Centrality-Total Degree		
Rank	Automap Run1	Automap Run2	AnalystOrg
1	UBL	UBL	UBL
2	abdul_aziz	abdul_aziz	Atta
3	nawaf_alhazmi	Atta	KSM
4	Atta	KSM	Atef
5	nawaf_al_hazmi	Ramzi	AHMED AL-NAMI

Table 7. ORA-based Top Five Leaders (Centrality Measures)in the Automap-assisted and Analyst Datasets.

POWER

The following tables are a comparison of very different network statistics. Clique Count, Cognitive Demand, & Knowledge Exclusivity indicate different types of power in the terrorist organization that results from both power within the agent x agent network and the attributes and resource/knowledge/ etc networks. In these measures, the datasets had the most divergence and an analyst using one or the other datasets may have come to very different conclusions about the threat.

Clique Count: Only considers the agent matrix. People that are highly effective tend to surround themselves with very tight knit subgroups, we call these cliques. Agents high in clique count are members of many of these small highly effective and integrated workgroups. All of the datasets agree and point to UBL, KSM, and Atta as linchpins in the organization in 2001.

Clique Count		
Automap Run1	Automap Run2	Analyst Org
UBL	UBL	UBL
KSM	abdul_aziz	Atta
abdul_aziz	KSM	KSM
Ramzi	Atta	Khalid al Mihdhar
Atta	Ramzi	Atef

Table 8. ORA-based Top Five Leaders (Clique Count) in the Automap-assisted and Analyst Datasets.

Cognitive Demand: Considers the agent, task, and knowledge matrices. A person high in cognitive demand is considered an emergent leader. Often overlooked, they may have fewer agent connections, but they make up for it in their knowledge and abilities. The emergent leaders from each dataset are significantly different. In large part this is because the Automap-assisted dataset has a much richer task and knowledge adjacency matrix. The identification of Ziad Jarrah in the Analyst generated dataset is striking. Current thinking is that there are two Ziad Jarrahs, one that flew flight 93 and another still on the loose. However, the analysts rolled all Ziad Jarrahs into a single individual

Cognitive Demand		
Automap Run1	Automap Run2	Analyst Org
abdul_aziz	fahd_al-quso	Ziad Jarrah
UBL	UBL	Atta
nawaf_al_hazmi	abdul_aziz	ZAKARYA ESSABAR
ali_ghamdi	Atta	MARWAN AL-SHEHHI
Atta	Ramzi	WADIH EL-HAGE

Table 9. ORA-based Top Five Leaders (Cognitive Demand) in the Automap-assisted and Analyst Datasets.

Knowledge Exclusivity: Considers the knowledge and agent matrix. There is an old adage that knowledge is power. In this case, these people have knowledge that few other organizational members possess. Depending on the content of the knowledge they can become critical to the organization without showing high on the agent matrix measures. Again, the paucity of knowledge data in the Intelligence Analyst approach gives limited credibility to the analysis. However, the rich data created by Automap identifies Abdelghani Mzoudi as a critical member of the 911 team. In fact, his connections to all of the 911 hijackers lead to a five month trial but, ultimately, acquittal.

Knowledge Exclusivity		
Automap Run1	Automap Run2	Analyst
zacarias_moussaoui	abdelghani_mzoudi	Ziad Jarrah
ali_ghamdi	khadr	JOSE PADILLA
abdul_aziz	tawfiq_attash	WADIH EL-HAGE
UBL	Ramzi	Atta
abu_zubaydah	janet_reno	SALEEM AL-HAZMI

Table 9. ORA-based Top Five Leaders (Centrality Measures)in the Automap-assisted and Analyst Datasets.

4. Discussion

Text-to-Data Process: The Automap-assisted approach is clearly less painful (resource intensive) than the pure human-cognitive approach to developing datasets. An analyst working with Automap took significantly less time and resources to complete the same tasks. In fact, an Automap-assisted analyst could have iteratively tweaked and analyzed the raw data more than nine times before the team of three human-cognitive process analysts had completed a single iteration. The Automap-assisted approach effectively lowers the threshold for redoing the coding and analysis from scratch as our intelligence community learns of new variables to encode datasets.

Database Product: In terms of database product, the Automap-assisted and Intelligence Analyst approaches were complementary. Both approaches yielded a significant number of unique agents that were not present in the other. Taking an Automap only approach would have missed Al Qaeda members such as Hamza al Ghamdi, Hasan Ghul, or Jose Padilla. While taking a Intelligence Analyst only approach would have missed Al Qaeda members and supporters such as Abu Al-Zarqawi, Baker Bashir, Mullah Omar, Richard Reid, or Zacarias Moussaoui.

One area where the Automap-assisted process requires further analysis is in edge or relationship detection. The human-cognitive process identified 50% more relationships between the nodes than the Automap-assisted process. In part this is accounted for in the significantly greater number of nodes found in the Automap-assisted process. However, it is also possible that the human-cognitive process finds implicit relationships that are not readily apparent to the Automap-assisted explicit processes.

But the biggest issue here is determining the nature of the edge…is it a relationship or just an insignificant interaction. In other words, the edge is established not just on the basis of one report but by a number of reports that suggest a repeated and sustained relationship. Or alternatively, a relationship is established based on the content of the data., semantic cues that an expert analyst, using informed intuition can then infer relationship over interaction.

Threat Analytic Product: An analyst armed with one or the other database products would have offered very different views of the organization and key leadership. The Automap-assisted data shows Al Qaeda as an organization with limited connections, few leaders and very diverse knowledge and objectives. The Intelligence Analyst data shows Al Qaeda as an organization with more connections, many leaders, and a narrow knowledge set and shared objectives.

Which threat analytic product is the better description is the million dollar question. However, the answer appears to be that, in combining the data from each approach, we have a more complete picture of the terrorist organization. For even with the significant difference in original database product, there were similarities (40-60% equivalent) and useful differences in the threat analytics of key terrorists. The Intelligence Analyst approach seems to see more subjective connections than there may actually be in the data and seems to preclude the inclusion of agents that are not known as terrorists. The Automap approach seems to see less connections than there may be between agents and seems to include many peripheral agents that may have nothing to do with Al Qaeda. It is possible that in combining the datasets we can balance the shortcomings of each approach.

Limitations

This research depended upon the use of novices in each of the conditions. The three analysts that performed the human unassisted database building effort were performing their first intelligence task after training. Further, the analyst assisted by Automap was using the tool for the first time and in a unfamiliar context (stratiegic counter-terrorism). It is likely that both methods would have benefited from experienced participants.

There was no ground truth to compare the results of the experimental conditions against. The ground truth network would consist of only the relevant nodes and the relevant relationships that could be explicitly and implicity determined from the texts. This is a problem common to this type research as full hindsight of all future terror activities would be required to determine ground truth.

As a result, this research was unable to identify false positives and false negatives. False positives, a node or edge added to the network that is not there, could result in the false arrest or worse of an innocent individual. False negatives, nodes or edges identified as not relevant to the network when they should be, could result in a key individual slipping through our dragnet and doing harm to the general population. Knowing which method, analyst or Automap, is succeptible to false positives and false negatives is critical to developing trust in the results.

5. Conclusion

There is an old adage that, "you make decisions based on the truth you have". No where is this more appropriate than in counter-terrorism analysis. Intelligence Analysts have to make their assessments based on the available truth yielded from the chosen entity and relationship extraction (ERE) process. As the search for WMD in Iraq has demonstrated, incomplete data can yield weakly formed analysis and ill-informed decisions [12].

When the available truth is further limited because intelligence reports from the field are left unprocessed the problem is further compounded [27]. Each unprocessed report has the ability to confirm or deny a node or edge in the terror network database. The more reports left unprocessed, the more likely an intelligence agency will make the wrong assessment. To combat this problem, the US is in its largest intelligence analyst hiring process for a decade[16]. However, without tools that assist the intelligence analyst's process it is unlikely our databases will catch up with the flow of field reports.

Automap, and similar text processing methods, offer a solution now. This study showed that a single analyst running the Automap software can process raw field reports in one tenth of the time of a team of analysts. However, the Automap-assisted approach has its weaknesses, namely the relatively unsophisticated window size methodology that establishes edges on the basis of a blanket proximity calculation. The resulting dataset may not represent all implicit data points, but, as this paper demonstrated, neither do the human intelligence analysts current subjective and intuition-based methods.

Human-centered automation has been one effective solution to similar problems. The history of the progress of man is about how he has leveraged technology to expand his limited cognitive capacities using tools - the microscope, the telescope and now in the modern age, tools that help him gain access to physical/knowledge ghosts whose presence cannot be doubted but can only be seen through the application of quantitative methods. With Automap, the intelligence analyst is still responsible for coding the most important field reports and culling through the implicit data points, while the tool insures that all field reports get into the database without backlog.

What can be said is that informed by both approaches the analyst gains an enhanced understanding of any given network. However the target folder subjective approach is suboptimal and the Automap approach remains a proof of concept with a clear agenda to enhance how the tool draws edges between identified nodes. Clearly however, the juxtaposition of both approaches highlight the importance of powerful tools to do what is likely the most difficult aspect of network analysis: data distillation, preparation and pre-processing. Through this combination of methods, our intelligence analyst will be closer to working from every field report our intelligence agents in the field are risking their lives to collect.

ACKNOWLEDGEMENTS
This work would not have been possible without Major Drew Cukor, USMC, Center for Strategic Studies. He agreed to share his experiences and data sets as he developed his own implementation of CASOS tools for the intelligence community.

6. References

[1]Army FM 34-36 (1991) Special Operations Forces Intelligence Operations. Headquarters, Department of the Army Washington, DC 30 September 1991.

[2]Carley, Kathleen M. (1986) "An Approach for Relating Social Structure to Cognitive Structure." Journal of Mathematical Sociology 12: 137-189.

[3]Carley, Kathleen M. (1988) "Formalizing the Social Expert's Knowledge." Sociological Methods and Research 17.2 (1988): 165-232.

[4]Carley, Kathleen M. (1993) "Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis." Sociological Methodology 23, Eds. Peter V. Marsden. Cambridge, MA: Blackwell Publishers. 75-126.

[5]Carley, Kathleen M. (1997a) "Network Text Analysis: the network position of concepts." Text analysis for the social sciences, Ed. Carl W. Roberts. Mahwah, NJ: Lawrence Erlbaum Associates, Inc. 1997a. 79-102.

[6]Carley, Kathleen M. (1997b) "Extracting team mental models through textual analysis." Journal of Organizational Behavior 18: 533-558.

[7]Carley, Kathleen M., and Michael Palmquist. (1992) "Extracting, Representing, and Analyzing Mental Models." Social Forces 70.3 :601-636.

[8]Carley, Kathleen M., "Dynamic Network Analysis." Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers, Eds. Ronald Breiger, Kathleen Carley, and Philippa Pattison, Committee on Human Factors, National Research Council, National Research Council. 2003. 133-145

[9]Carley, Kathleen M., and David Krackhardt. (1998) "A PCANS Model of Structure in Organization." Proceedings of the 1998 International Symposium on Command and Control Research and Technology. Monterray, CA. Evidence Based Research, Vienna, VA. 113-119.

[10]Carley, Kathleen M., and Yuqing Ren. (2001) "Tradeoffs Between Performance and Adaptability for C3I Architectures." Proceedings of the 2001 Command and Control Research and Technology Symposium. Annapolis, Maryland. Evidence Based Research, Vienna..

[11]Carley, Kathleen & Reminga, Jeffrey. (2004). ORA: Organization Risk Analyzer. www.casos.cs.cmu.edu

[12]Clarke, Richard (2004) Against All Enemies: Inside America's War on Terror, Simon & Schuster, Free Press.

[13]Department of Justice (2003). Law Enforcement Intelligence: A Guide for State, Local, and Tribal Law Enforcement Agencies. www.cops.usdoj.gov

[14]Diesner, Jana, Eleanor T. Lewis, and Kathleen M. Carley. (2003)"How You Code Matters: How Coding Techniques and Choices Influence Results from Automated Text Analysis", CASOS working paper..

[15]Diesner, Jana & Carley, Kathleen. (2004). Using Network Text Analysis to Detect the Organizational Structure of Covert Networks.. In Proc of NAACSOS. Pittsburgh, PA

[16]Ensor, David (2005) "Cloak, dagger and résumé", www.cnn.com

[17]Gerner, D.J., Schrodt, P.A., Francisco, R., & Weddle, J. L. (1994). The analysis of political events using machine coded data. International Studies Quarterly 38,91-119.

[18]I2 Inc. (2004). http://www.i2inc.com/

[19]Kean, Thomas-Chair (2004). The 9/11 Commission Report: Final Report of the National Commission on Terrorist Attacks Upon the United States by National Commission on Terrorist Attacks. W.W. Norton, New York

[20]Kovar, K., J. Fürnkranz, J. Petrak, B. Pfahringer, R. Trappl, and G. Widmer. 2000. Searching for Patterns in Political Even Sequences: Experiments with the KEDS Database. Cybernetics and Systems 31, 6.

[21]Lormel, D.M. (2002). Testimony of Dennis M. Lormel, Chief, Terrorist Financial Review Group, FBI Before the Senate Judiciary Committee Subcommittee on Technology, Terrorism and Government Information July 9, 2002 Hearing On S. 2541, "The Identity Theft Penalty Enhancement Act"

[22]Manno, C. (2004). Seventh public hearing of the National Commission on Terrorist Attacks Upon the United States. Statement of Claudio Manno to the National Commission on Terrorist Attacks Upon The United States January 27, 2004

[23]Ridder, J.A. de (1988). Computer-aided Evaluative Text Analysis (CETA 1.2). Software.Muiderberg: Coutinho.

[24]Ryan, Gery W., and H. Russell Bernard. "Data Management and Analysis Methods". Handbook of Qualitative Research, 2nd ed., Eds. Norman Denzin, and Yvonna Lincoln. Thousand Oaks, CA: Sage Publications. 2002. 769- 802.

[25]Sageman, Marc Understanding Terror Networks (Philadelphia, PA: University of Pennsylvania Press, 2004).

[26]Sarawagi, Sunita & Bhamidipaty, Anuradha. (2002) Interactive Deduplication Using Active Learning, In Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2002), Edmonton, Canada, July.

[27]Treverton, Gregory F & Wilson, Peter A. (2005) True Intelligence Reform Is Cultural, Not Just Organizational Chart Shift Christian Science Monitor, January 13.

[28]Woods, D.D. & Roth, E.M., (1988) Cognitive Systems Engineering. In M. Helander, editor, Handbook of Human-Computer Interaction, North-Holland, New York.