Chapter 5

INTELLIGENCE AND WARNING

Chapter Overview
Although terrorism depends on surprise, terrorist attacks are not random but require careful planning, preparation, and cooperation before execution. To avoid being preempted by authorities, terrorists may disguise their true identities or hide their activities with their illegal objectives and intents behind other legal activities. Similarly, criminals may try to minimize the possibility of being identified and captured by using falsified identities. To detect the hidden intents and potential for future attacks or offenses is the main goal of intelligence and warning systems. In this section we present four case studies addressing the intelligence and warning needs.
5.1 Case Study 1: Detecting Deceptive Criminal Identities

It is a common practice for criminals to lie about the particulars of their identity, such as name, date of birth, address, and social security number, in order to deceive a police investigator. The ability to validate identity can be used as a warning mechanism as the deception signals the intent to commit future offenses. In this case study we focus on uncovering patterns of criminal identity deception based on actual criminal records and suggest an algorithmic approach to revealing deceptive identities (Wang et al., 2004a).

Data used in this study were authoritative criminal identity records obtained from the Tucson Police Department (TPD). These records were structured database entries containing criminal identity information, such as name, date of birth (DOB), address, identification number (e.g., social security number), race, weight, and height. Over 1.3 million criminal identity records were stored in the TPD databases. In order to study the patterns of criminal identity deception, we selected from the TPD database 372 records involving 24 criminals -- each having one real identity record and several deceptive records. Guided by a veteran police detective with over 30 years of service in law enforcement, we carefully examined these 372 records and found that deception mostly occurred in specific attributes: name, address, birth date, and ID number. The identity deception patterns in this dataset are shown in Figure 5-1.

Name deception, occurring in most deceptive cases, includes giving a false first name and a true last name or vice versa, changing the middle initial, giving a name pronounced similarly but spelled differently, etc. Deception on DOB can consist of, for example, switching places between the month of birth and the day of birth. Similarly, ID deception is often made by changing a few digits of a social security number or by switching their places. In residency deception, criminals usually change only one portion of the address. For example, we found that in about 87% of cases criminals provided a false street number along with the true street direction, street name, or street type.

To automatically detect deceptive identity records, we employed a similarity-based association mining method to extract associated (similar) record pairs. Based on the deception patterns found we selected four attributes, name, DOB, SSN, and address, for our analysis. We compared and calculated the similarity between the values of corresponding attributes of each pair of records. If two records were significantly similar we assumed that at least one of these two records was deceptive.

Because the selected four attributes have mostly string values, we compared two attribute values based on their edit distance (Levenshtein, 1966) and Soundex code (Newcombe et al., 1959). The edit distance
between two strings is the minimum number of single character insertions, deletions, and substitutions required to transform one string into the other. Soundex code represents the phonetic pattern of a string. For example, “PEARSE” and “PIERCE” are both coded as “P620.” To detect both spelling and phonetic variations between two name strings, edit distance similarity and Soundex similarity were computed separately. In order to capture name exchange deception, similarities were also computed based on different sequences of first name and last name. We took the similarity value from the sequence that had the maximal value between two names. We used only edit distance to compare non-phonetic attributes of DOB, SSN, and address. Each similarity value was normalized between 0 and 1. The similarity value over all four attributes was calculated using a normalized Euclidean distance function.

In order to test the performance of our approach, another sample containing 120 criminal records with identified deception was chosen from the TPD database. We ignored records with missing values. The 120 records involved 44 criminals, each of whom had an average of three records in the sample set. Some data was used to train and test our algorithm so that records pointing to the same suspect could be associated with each other. Training and testing were validated by a standard hold-out sampling method.

Figure 5-1. Identity deception patterns. Each percentage number represents the proportion of records that contain the particular type of deception in the selected dataset.
Of the 120 records in the testbed, 80 (2/3) were used for training the algorithm, while the remaining 40 were used for testing purposes.

A similarity matrix was built for all training records. Using the similarity values in the matrix, threshold values were searched to distinguish between the similar pairs of records and the dissimilar pairs. Accuracy rates for correctly recognized similar pairs of records using different threshold values are shown in Table 5-1. When the threshold similarity value was set to 0.52, our algorithm achieved its highest accuracy of 97.4%, with relatively small false negative and false positive rates, both of which were 2.6%.

A similarity matrix was also built for the 40 testing records. By applying the optimal threshold value to the testing similarity matrix, records having a similarity value of more than 0.52 were considered to be pointing to the same offender and were associated together. The accuracy of association in the testing dataset is shown in Table 5-2. The result shows that the algorithm is effective (with an accuracy level of 94%) in linking deceptive records pointing to the same offender.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Accuracy</th>
<th>False Negative *</th>
<th>False Positive **</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.60</td>
<td>76.60%</td>
<td>23.40%</td>
<td>0.00%</td>
</tr>
<tr>
<td>0.55</td>
<td>92.20%</td>
<td>7.80%</td>
<td>0.00%</td>
</tr>
<tr>
<td>0.54</td>
<td>93.50%</td>
<td>6.50%</td>
<td>2.60%</td>
</tr>
<tr>
<td>0.53</td>
<td>96.10%</td>
<td>3.90%</td>
<td>2.60%</td>
</tr>
<tr>
<td>0.52</td>
<td>97.40%</td>
<td>2.60%</td>
<td>2.60%</td>
</tr>
<tr>
<td>0.51</td>
<td>97.40%</td>
<td>2.60%</td>
<td>6.50%</td>
</tr>
<tr>
<td>0.50</td>
<td>97.40%</td>
<td>2.60%</td>
<td>11.70%</td>
</tr>
</tbody>
</table>

* False negative: consider dissimilar records as similar ones.
** False positive: consider similar records as dissimilar ones.

Table 5-2. The accuracy of association in the testing data set.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Accuracy</th>
<th>False Negative</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.52</td>
<td>94.0%</td>
<td>6.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

While the above case study showed promising results, much more research is needed for deception detection, which we believe is a unique and critical problem for ISI.

5.2 Case Study 2: The Dark Web Portal

Because the Internet has become a global platform for anyone to disseminate and communicate information, terrorists also take advantage of the freedom of cyberspace and construct their own web sites to propagate terrorism beliefs, share information, and recruit new members. Web sites of
terrorist organizations may also connect to one another through hyperlinks, forming a “dark web.” We are building an intelligent web portal, called Dark Web Portal, to help terrorism researchers collect, access, analyze, and understand terrorist groups (Chen et al., 2004c; Reid et al., 2004). This project consists of three major components: Dark Web testbed building, Dark Web link analysis, and Dark Web Portal building.

5.2.1. Dark Web Testbed Building

Relying on reliable governmental sources such as the Anti-Defamation League (ADL), FBI, and United States Committee for a Free Lebanon (USCFL), we identified 224 U.S. domestic terrorist groups and 440 international terrorist groups. For U.S. domestic groups, group-generated URLs can be found in FBI reports and the Google Directory. For international groups, we used the group names as queries to search major search engines such as Google and manually identified the group-created URLs from the result lists. To ensure that our testbed covered major regions in the world, we sought the assistance of language experts in English, Arabic, and Spanish to help us collect URLs in several major regions. All URLs collected were manually checked by experts to make sure that they were created by terrorist groups. After the URL of a group was identified, we used the SpidersRUs toolkit, a multilingual digital library building tool developed by our lab, to collect all the web pages under that URL and store them in our testbed. Table 5-3 shows a summary of web pages collected from three rounds of spidering (performed bi-monthly).

5.2.2. Dark Web Link Analysis and Visualization

Terrorist groups are not atomized individuals but actors linked to each other through complex networks of direct or mediated exchanges. Identifying how relationships between groups are formed and dissolved in the terrorist group network would enable us to decipher the social milieu and communication channels among terrorist groups across different jurisdictions. Previous studies have shown that the link structure of the web represents a considerable amount of latent human annotation (Gibson et al., 1998). Thus, by analyzing and visualizing hyperlink structures between terrorist-generated web sites and their content, we could discover the structure and organization of terrorist group networks, capture network dynamics, and understand their emerging activities.
Table 5-3. Summary of URLs identified and web pages collected.

<table>
<thead>
<tr>
<th>Region</th>
<th>U.S.A. Domestic</th>
<th>Latin-America</th>
<th>Middle-East</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
<td>2nd</td>
<td>3rd</td>
</tr>
<tr>
<td>Batch #</td>
<td>Total</td>
<td>81</td>
<td>233</td>
</tr>
<tr>
<td># of seed URLs</td>
<td>From literature &amp; reports</td>
<td>63</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td>From search engines</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>From link extraction</td>
<td>18</td>
<td>120</td>
</tr>
<tr>
<td># of terrorist groups searched</td>
<td>74</td>
<td>219</td>
<td>71</td>
</tr>
<tr>
<td># of web pages</td>
<td>Total</td>
<td>125,610</td>
<td>396,105</td>
</tr>
<tr>
<td># of multimedia files</td>
<td>0</td>
<td>70,832</td>
<td>223,319</td>
</tr>
</tbody>
</table>

5.2.3. Dark Web Portal Building

Using the Dark Web Portal, users are able to quickly locate specific dark web information in the testbed through keyword search. To address the information overload problem, the Dark Web Portal is designed with post-retrieval components. A modified version of a text summarizer called TXTRACTOR, which uses sentence-selection heuristics to rank and select important text segments (McDonald and Chen, 2002), has been added into the Dark Web Portal. The summarizer can flexibly summarize web pages using three or five sentences, so that users can quickly get the main idea of a web page without having to read through it. A categorizer organizes the search results into various folders labeled by the key phrases extracted by the Arizona Noun Phraser (AZNP) (Tolle and Chen, 2000) from the page summaries or titles, thereby facilitating the understanding of different groups of web pages. A visualizer clusters web pages into colored regions using the Kohonen self-organizing map (SOM) algorithm (Kohonen, 1995), thus reducing the information overload problem when a large number of search results are obtained. Post-retrieval analysis could help reduce the information overload problem. However, without addressing the language barrier problem, researchers are limited to the data in their native languages and cannot fully utilize the multilingual information in our testbed. To address this problem, we added a cross-lingual information retrieval (CLIR) component into the portal. Based on our previous research, we have
developed a dictionary-based CLIR system for use in the Dark Web Portal. It currently accepts English queries and retrieves documents in English, Spanish, Chinese, and Arabic. Another component that will be added to the Dark Web Portal is a machine translation (MT) component, which will translate the multilingual information retrieved by the CLIR component back into the users’ native languages.

We show a sample search session in the figures below. Suppose the user is interested in the terrorist group “Ku Klux Klan” and uses it as a search term. Two types of search forms are available: simple search and advanced search (see Figure 5-2). Our user chose to use the simple search first. The advanced mode gives users more options to refine their search. For example, the user can specify web pages with the exact phrase. In addition, the user can restrict the results within a few terrorist categories or choose to search a particular file type, such as PDF or Word files.

![Figure 5-2. Dark Web Portal interfaces: simple search and advanced search.](image)

![Figure 5-2a. U.S. domestic (English) simple search interface.](image)
By hitting the “Find Results” button, the top 20 results are displayed (see Figure 5-3). On the top of the result page it shows a list of “suggested keywords,” such as “Aryan Nations” and “David Duke,” which helps the user to expand or refine the query. Along with the web page result display, our portal also presents the terrorist group name and the corresponding group category. As terrorist group web pages may often disappear, “Cached Pages” for each web page collected at different time periods are provided (e.g., 2004/03). Additionally, the user can view web pages, PDF files, or Word files by clicking the corresponding links.

As terrorist groups continue to use the Internet as their communication, recruiting, and propaganda tool, a systematic and system-aided approach to studying their presence on the web is critically needed.
5.3 Case Study 3: Jihad on the Web

With weekly news coverage of excerpts from videos produced and web cast by terrorists, it has become clear that terrorists have further exploited the Internet beyond routine communication and propaganda operations to better influence the outside world (Arquilla & Rondeldt, 1996). Some terrorism researchers posited that terrorists have used the Internet as a broadcast platform for the “terrorist news network,” which is an effective tactic because they can reach a broad audience with relatively little chance of detection (Elison, 2000; Tsfati & Weimann, 2002; Weinmann, 2004). Although this alternate side of the Internet, referred to as the “Dark Web,” has recently received extensive government and media attention, systematic understanding of how terrorists use the Internet for their campaign of terror is very limited.

In this study, we explore an integrated computer-based approach to harvesting and analyzing web sites produced or maintained by Islamic Jihad extremist groups or their sympathizers to deepen our understanding of how
Intelligence and Warning

Jihad terrorists use the Internet, especially the World Wide Web, in their terror campaigns. More specifically, we built a high-quality Jihad terrorism web collection using a web harvesting approach and conducted hyperlink analysis on this collection to reveal various facets of Jihad terrorism web usage. We hope to supplement existing high-quality but manually-driven terrorism research with a systematic, automated web spidering and mining methodology.

5.3.1. Building the Jihad Web Collection

To guarantee that our collection is comprehensive and representative, we take a three-step systematic approach to construct our collection:

1. Identifying seed URLs and backlink expansion: The first task is to find a small set of high-quality Jihad web sites. To identify terrorist groups, we completely relied on the U.S. Department of State’s list of foreign terrorist organizations. In particular, we only selected Middle-Eastern organizations from that list for this study. After identifying the terrorist groups in the Middle-East region, we manually searched major search engines to find the web sites of these groups. Our goal was not to construct a comprehensive list of URLs, but merely to compile a small list of high-quality URLs that can serve as the seeds for backlink expansion. The backlinks of these URLs were automatically identified through Google and Yahoo backline search services and a collection of 88 web sites was automatically retrieved.

2. Manual collection filtering: Because bogus or unrelated terrorist sites can make their way into our collection, we developed a manual filtering process based on evidence and clues in the web sites. Aside from sites which explicitly identify themselves as the official sites of a terrorist organization or one of its members, a web site that contains praise of or adopts ideologies espoused by a terrorist group is included in our collection.

3. Extending search: To ensure the comprehensiveness of our collection we augment the collection by means of expanded searches. Based on the 26 web sites identified in the previous step, we constructed a small lexicon of commonly-used Jihad terms with the help of Arabic language speakers. Examples of highly relevant keywords included in the lexicon are: "الكفار" ("Infidels"), "حرّب صليبية" ("Crusader’s War"), "المجاهدين" ("Moujahedins"), etc. This lexicon is utilized to perform expanded searches. The same rules used in the filtering process are used here to discern fake and unrelated web sites. As a result, our final Jihad web collection
contains 109,477 Jihad web documents including HTML pages, plain text files, PDF documents, and Microsoft Word documents.

5.3.2. Hyperlink Analysis on the Jihad Web Collection

We believe the exploration of hidden Jihad web communities can give insight into the nature of real-world relationships and communication channels between terrorist groups (Weimann, 2004). Uncovering hidden web communities involves calculating a similarity measure between all pairs of web sites in our collection. We define similarity as a function of the number of hyperlinks in web site “A” that point to web site “B,” and vice versa. In addition, a hyperlink is weighted proportionally to how deep it appears in the web site hierarchy. The similarity matrix is then used as input to a Multi-Dimensional Scaling (MDS) algorithm (Torgerson, 1952), which generates a two dimensional graph of the web sites. The proximity of nodes in the graph reflects their similarity level.

As shown in Figure 5-4, domain experts recognized six clusters representing hyperlinked communities in the network. On the left side of the network resides the Hizbollah cluster. Hizbollah is a Lebanese militant organization. Established in 1982 during the Israeli invasion of Lebanon, the group routinely attacked Israeli military personnel until their pullout from south Lebanon in 2000. A cluster of web sites of Palestinian organizations occupies the bottom-left corner of the network, including: Hamas, Al-Aqsa Martyr’s Brigades, and the Palestinian Islamic Jihad. An interesting observation here is the close link between the Hizbollah community and the Palestinian militant groups’ community. Hizbollah has traditionally sympathized with the Palestinian cause.

On the top-left corner sits the Hizb-ut-Tahrir cluster. Hizb-ut-Tahrir is a political party with branches in many countries in the Middle-East and in Europe. Although the group is believed to be associated with Al-Qaeda, an apparent relationship between the two groups has not been proven. Looking at the bottom-right corner, one can see a cluster of Al-Qaeda affiliated sites. This cluster has links to two radical Palestinian web sites. Al-Qaeda sympathizes with Palestinian groups and some Palestinian Islamist groups like Hamas and Islamic Jihad share the same Salafi ideology with Al-Qaeda. In the top-right corner, the Jihad Sympathizers web community includes web sites maintained by sympathizers of the Global Salafi movement. For example, “kavkazcenter.net” and “clearguidance.com” are two web sites maintained by sympathizers of the Chechen rebels. As expected, the sympathizers’ community does not have any links to Hizbollah’s community as they follow radically different ideologies.
Figure 5-4. The Jihad terrorism web site network visualized based on hyperlinks.

Visualizing hyperlinked communities can lead to a better understanding of the Jihad web presence. Furthermore, it helps foretell likely relationships between terrorist groups.

5.4 Case Study 4: Analyzing the Al-Qaeda Network

As part of the worldwide Islamic revivalist movement, a number of terrorist organizations have targeted the West. Terrorism and terrorist attacks pose severe threats and have brought significant damage to the whole world. Only with an in-depth understanding of terrorism and terrorist organizations can we defend against the threats. Because terrorist organizations often operate in a network form in which individual terrorists cooperate and collaborate with each other to carry out attacks (Klerks, 2001; Krebs, 2001), network analysis methodology can help discover valuable knowledge about terrorist organizations by studying the structural properties of the networks (Xu and Chen, Forthcoming). We have employed techniques and methods
from social network analysis (SNA) and web mining to address the problem of structural analysis of terrorist networks.

The objective of this case study is to examine the potential of network analysis methodology for terrorist analysis. By comparing our findings with experts’ input we would be able to ascertain whether automatic analysis of structural properties of a terrorist network would generate valuable knowledge that is consistent with expert knowledge. In this study, we focus on the structural properties of a set of Islamic terrorist networks including Osama bin Laden’s Al-Qaeda. A recently published book (Sageman, 2004) documents the history and evolution of these terrorist organizations, which are called Global Salafi Jihad (GSJ) in this book. The author of this book is a social psychologist and used to be a Foreign Service officer. During the Afghan-Soviet war from 1986 to 1989, he dealt with Islamic Fundamentalists on a daily basis and acquired a substantial amount of expertise in terrorism and terrorist organizations.

Based on various open sources such as news articles and court transcripts, the author collected data about 364 terrorists in the GSJ network regarding their background, religious beliefs, social relations, and terrorist attacks they participated in. There are three types of social relations among these terrorists: personal links (e.g., acquaintance, friendship, and kinship), operational links (e.g., collaborators in the same attack), and relations formed after attacks (Sageman, 2004). The author identified within this network four major terrorist groups based on their geographical locations: Central Staff, Core Arab, Maghreb Arab, and Southeast Asian. Each group has its own leaders. For example, Osama bin Laden is the leader of the Core Staff group, which connects to the other three groups through several lieutenants.

We analyzed the GSJ network based on the social relation data contained in a spreadsheet provided by the author. Using the SNA visualization approach we depicted the GSJ network graphically as shown in Figure 5-5.

- **Centrality analysis.** Considering all three types of social relations, we found that the four group leaders were among the top 11 most popular members, where the popularity was represented by degree measure. For example, Osama bin Laden had 72 links to other terrorists and ranked the second in degree. Although he was not a leader, Hambali had the highest degree score and played an important role in connecting different terrorist groups (see Figure 5-5a). Moreover, the lieutenants tended to have high scores in betweenness and served as gatekeepers between groups. The analysis implies that centrality measures could be useful for identifying important members in a terrorist network.
Subgroup analysis. The four terrorist groups are shown in Figure 5-5b, based on the author’s input. To find out whether these geographical-based groups were also cohesive groups in structure, we calculated the cohesion score (Wasserman and Faust, 1994) of each group. We found that all these groups had high cohesion scores. The Southeast Asian group scored the highest in cohesion. This may suggest that members in this group tended to be more closely related to members in their own group than to members from other groups. According to the author, the Southeast Asian group was quite different from the other three groups in terms of their religious beliefs and missions.

Network structure analysis. According to the author, these groups had different structures: the Southeast Asian group’s structure was hierarchical, in which members at higher levels lead lower-level members, while the other three groups were scale-free networks (Albert and Barabasi, 2002). However, we found that the four groups were similar in their degree distribution, which was a power-law distribution with a long tail for large values of degree. This implies that all four networks were scale-free networks in which a few important members (nodes with high degree scores) dominated the network and new members tend to join a network through these dominating members. This finding has an important policy implication, that is, disruptive strategies should potentially be focused on central members in a terrorist network.

Link path analysis. Comparing the personal network representation (Figure 5-5b) and the operational network representation (Figure 5-5c), we found that some important members did not have direct personal links to an attack prior to the execution of the attack. For example, Osama bin Laden, KSM, and Hambali had no direct personal links to terrorists in the 9/11 attack clique. We performed link path analysis to find out the shortest paths of personal links that lead to 9/11 terrorists. One of our hypotheses was that Osama bin Laden connected to the 9/11 clique through a four-hop path: bin Laden – Nashiri – ZaMihd – Mihdhar – Shibh (the highlighted path in Figure 5-5b). Although this hypothesis turned out to be wrong based on the author’s feedback (other information was needed to establish the link), the analysis showed the potential of using link path analysis to generate hypotheses about the motives and planning processes of terrorist attacks.
Figure 5-5. The Global Salafi Jihad (GSJ) network.
Figure 5-5a. The GSJ network with all types of relations. Each node represents a terrorist. A link represents a social relation. The four terrorist groups are color-coded in original figures (not shown here): Central Staff—pink, Core Arab—yellow, Maghreb Arab—blue, and Southeast Asian—green. Leaders are labeled in red and lieutenants are labeled in black.
Figure 5-5b. The GSJ network with personal links. The highlighted path indicates the hypothesis regarding the connection between bin Laden and the 9/11 attacks.
5.5 Future Directions

Deception detection is one of the unique problems facing law enforcement and intelligence applications. Criminals and terrorists often try to hide or disguise their identities via various means. Biometric (e.g., fingerprints and DNA) techniques and behavioral or psychological interrogations have been widely adopted in law enforcement and intelligence communities. With the abundance of criminal identity information available in law enforcement and intelligence databases and the ability of advanced algorithms for “fuzzy” queries and entity matching, we believe data mining guided deception detection research shows tremendous potential.

We also believe that web-based open source “terrorism informatics” research is critically needed for terrorism research and intelligence analysis purposes. There is a pressing need for traditional terrorism researchers and
analysts to leverage the new advances in web retrieval, mining, analysis, and visualization. By combining the domain expertise and methodology well established in terrorism research with new information technologies, we believe the new science of terrorism informatics would emerge and contribute to the systematic study and understanding of the global terrorism phenomena.

5.6 Questions for Discussion

1. What are other deception detection applications in the law enforcement or intelligence community? How can relevant datasets be obtained?

2. What are some ways to identify terrorism researchers? What are the major terrorism research conferences and publications? What are the major terrorism research centers?

3. What are some ways to collaborate with the intelligence community for non-security clearance level research? What are some of the rich open sources for technical ISI research?

4. What are some ways to develop and advance multilingual and multimedia research in ISI? What are the technical foundations and promising approaches?