A Survey on Terrorist Network Mining: Current Trends and Opportunities

Nisha Chaurasia[1], Mradul Dhakar[1], Akhilesh Tiwari[2] and R. K. Gupta[2]

[1]MTech Student, Department of CSE & IT, MITS Gwalior, M.P., India
chaurasianisha21@gmail.com, mraduliitm@gmail.com
[2] Department of CSE & IT, MITS Gwalior, M.P., India
atiwari.mits@gmail.com, iiitm_rkg@rediffmail.com

ABSTRACT

Along with the modernization and widespread usage of Internet, the security of the mankind has become one of the major issues today. The threat of human society from the terrorists is the challenge faced dominantly. Advancement in the technology has not only helped the common people for the growth but also these inhuman people to adversely affect the society with sophisticated techniques. In this regard, the law-enforcement agencies are aiming to prevent future attacks. To do so, the terrorist networks are being analyzed and detected. To achieve this, the law enforcement agencies are using data mining techniques as one of the effective solution. One such technique of data mining is Social network analysis which studies terrorist networks for the identification of relationships and associations that may exist between terrorist nodes. Terrorist activities can also be detected by means of analyzing Web traffic content. This paper studies social network analysis, web traffic content and explores various ways for identifying terrorist activities.

KEYWORDS

Associations, Data mining, Social network analysis, Web traffic content.

1. INTRODUCTION

The concern about the global security came into limelight after 9/11 attacks. A major challenge faced by the law enforcement agencies is the large crime 'raw' data volumes and the lack of sophisticated network tools and techniques to utilize the data effectively and efficiently. Likewise, the web traffic generates a vast amount of data from which only a small portion is critical to the intelligence.

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information-information that can be used to increase revenue, cut costs, or both [4]. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified [2].

Terrorist network mining has emerged as a novel field of research often applied to investigation of organized crimes. Relationship among criminals/terrorists form the basis for the organized crimes and are essential for smooth operation of a criminal/terrorist organization which can be

viewed as a network where nodes represents terrorists and links represent relationships or associations between terrorists [6].
Traditionally, analysis of terrorist network was a manual process consuming much time and effort due to information overload and thus failed to generate valuable knowledge on time hence effective techniques are of essence to amend the information overload problem. This paper describes the techniques that generate patterns distinguishing between legitimate and threat groups and helps law enforcement agencies to decide which networks to put under scrutiny.
Rest of the paper is organized as follows: Section 2 discusses about the analysis of terrorist network using SNA (Social Network Analysis) technique. Section 3 explains detection of terrorist networks by applying content-based detection methodology. Section 4 offers the ways for destabilization of terrorist networks. Further, Section 5 talks about the future work and conclusion.

2. SOCIAL NETWORK ANALYSIS

Social Network Analysis (SNA) is one of the most preferred technologies for studying criminal and terrorist networks. The SNA technique defines the roles and interaction among the actors within the social network. The social network is a social structure consists of individuals, sometimes covert, human groups, and organizations and defines relationships such as friendship, kinship among them. These interrelationships are represented through graph using SNA. The graph is build by analyzing the data comprising of nodes (terrorists) and links (relationships). Investigators use SNA to analyze these relationships to deduce information about the individuals and groups.
Valdis Krebs encountered three problems while analyzing 9/11 hijacking network. These are: [8]

- INCOMPLETENESS: the covert graph is incomplete due to missing actors (nodes) and links (edges) that investigators may fail to uncover.
- FUZZY BOUNDARY: the difficulty in deciding who to include and who not to include.
- DYNAMIC: terrorists network are dynamic as actors are added or removed frequently.

In addition, the emphasis is also laid on measuring strength of ties between actors in terrorist networks. The strong ties depict the relationship with family or close friend while the weak ties depict relationship with acquaintances.
Specifically, SNA is capable of detecting subgroups discovering their patterns of interaction, identifying central individuals.

2.1 Subgroup Detection

SNA uses the cluster analysis to partition the network into subgroups of individuals who interact with each other called clusters which are not otherwise apparent in data.

2.2 Discovery of pattern of interaction

The interaction between the subgroups or clusters is discovered using SNA approach called Block modeling [7]. The block modeling approach can uncover the patterns of between-group

interactions and associations. The approach defines the presence or the absence of an association between the subgroups by measuring the link density, calculated as:

$$\rho_{ij} = L_{ij}/n_i * n_j$$

Where, i and j are the two subgroups. L_{ij} is the actual number of links between subgroups. n_i and n_j are the number of nodes with subgroups i and j respectively.
This calculated ρ_{ij} is compared against a predefined threshold value. If the ρ_{ij} is greater than the threshold value then an association between subgroups is present. Thus have a strong association indicating that the two subgroups are interacting with each other constantly.

2.3 Centrality measures

The network is represented by a simple and undirected graph. Mathematically, a network is represented by the adjacency matrix A as:

$$\begin{cases} A_{ij} = 1 & \text{if i and j are connected} \\ 0 & \text{otherwise} \end{cases}$$

The matrix is a symmetric matrix i.e. $A_{ij} = A_{ji.}$

The importance of each network member is identified using centrality measure from SNA. Centrality measures helps in identifying the key player or central person in the network. Several centrality measures are defined such as degree, betweenness, closeness, and eigenvector and can suggest the importance of node in a network.
The *degree* of a node is its number of links. An individual's having a high degree, for instance, may imply his leadership. The degree D_i of a vertex i is [6]:

$$D_i = \sum_{j=1}^{n} A_{ij}$$

Betweenness is the number of geodesics (shortest paths between any two nodes) passing through it. An individual with high betweenness may be a gatekeeper in the network. It is given by [6]:

$$B_a = \sum_{a}^{n} \sum_{j}^{n} g_{ij}(a)$$

Where g_{ij} (a) indicates whether the shortest path between two other nodes i and j passes through node a.
Closeness is the sum of all geodesics between the particular node and every other node in the network [6]:

$$C_a = \sum_{i=1}^{n} l(i, a)$$

Where l (i, a) is the length of the shortest path connecting nodes i and a.
Eigenvector centrality acknowledges that not all connections are equal. If we denote the centrality of vertex *i* by x_i, then we can allow for this effect by making x_i proportional to the average of the centralities of *i*'s network neighbors [6].

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j$$

Where λ is constant. Defining the vector of centralities x = (xl; x2; : : :), we can rewrite this equation in matrix form as:

$$\lambda x = A * x$$

Hence we see that x is an eigenvector of the adjacency matrix with eigenvalue λ.
Baker and Faulkner employed these four measures, especially degree to find the central individuals in a price-fixing conspiracy network in the electrical equipment industry [6].

3.CONTENT-BASED TERRORIST DETECTION METHODOLOGY

In the content-based detection methodology, the terrorist are detected by the Web traffic content by monitoring all ISPs traffic. The prior knowledge about the terrorists is maintained as the content for training the detection algorithm. This detection should be carried out in real time.

3.1 Intrusion Detection System

The detection of the content from the existing sites and known terrorist traffic on the web is done by using Intrusion Detection System (IDS).The IDS constantly monitors the various activities within the network traffic to estimate the possible hostile attacks. The IDS could be a computer, or computers within the network or the network itself. The IDS analyzes the various activities information gained and evaluates the possibility of the intruders within the network. The measures evaluated by the IDS include accuracy, completeness, performance, efficiency, fault tolerance, timeliness and adaptivity.

3.2 Vector-space Model

The information evaluated by the IDS is represented in the textual content of Web pages. The document D is represented by the n-dimensional vector V= (v_1, v_2, v_3…$v_{n)}$ where v_i represents the frequency-based weight of term i in D.
The similarity between the two documents represented as vectors may be computed by using one of the known vector distance measuring methods such as Euclidian distance or Cosine. The cosine similarity measure is commonly used to estimate the similarity between an accessed Web page and a given set of terrorists' topic of interest [9].

3.3 Clustering Techniques

In the clustering technique, we perform cluster analysis-process of partitioning data objects (records, documents etc) into meaning groups or clusters so that the objects within the cluster have similar characteristics but are dissimilar to objects (data) in other clusters [5].
The unsupervised clustering is performed for classification of patterns. The clustering is performed on the Web documents by clustering them into documents of similar interest. A centroid is calculated for each collection of cluster and represented by vector space model.

3.4 Content-based Detection of Terror-Related Activities

The detection of the terrorist is done through the content of the web pages browsed by the terrorist and their supporters. These Web pages content refers only to the textual content performing learning process to obtain 'Typical-Terrorist-Behavior'-defined as an access to information relevant to the terrorist and their supporters.
This methodology has two modes of operation:

3.4.1 Learning typical terrorist behavior

In the learning mode, the Web pages are downloaded from the terrorist related sites. The downloaded Web pages are clustered by applying unsupervised clustering techniques. Cluster serves as the data indicating the typical-terrorist behavior or the profile of the terrorists and their supporters. The Web pages downloaded are fed as the input to the Vector Generator module as shown in Figure 1, which converts the pages into vectors of weighted terms (each page into a vector). These vectors are stored for future processing in the vector of Terrorist Transaction in DB [9].
The unsupervised clustering is performed on these vectors. A centroid vector (C_i) is computed for each cluster by the Terrorist-Representor module and thus representing the typical-terrorist behavior.

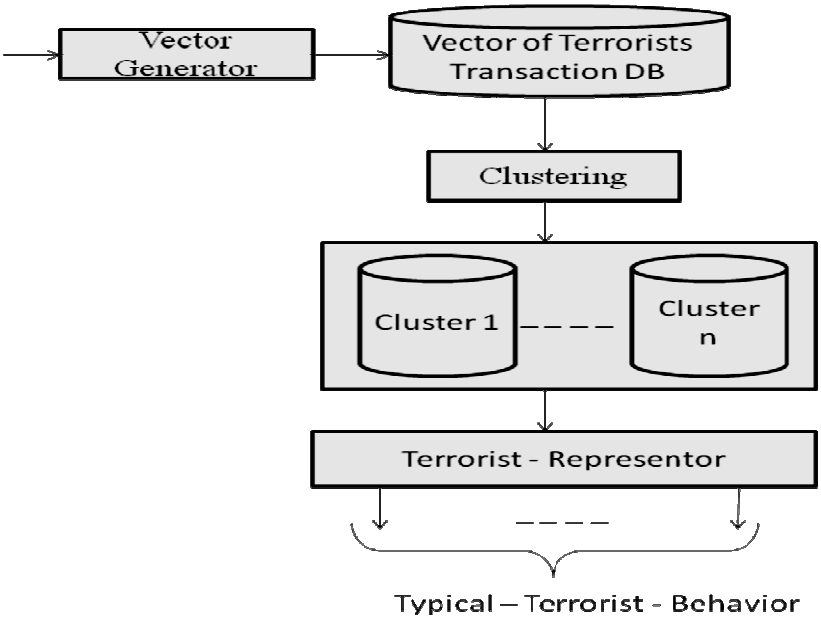

Figure 1: Learning the Typical-Terrorist-Behavior

3.4.2 Monitoring users

In this mode, the comparison of the content of information accessed by the users and the typical-terrorist-behavior is made. The textual content is represented in the form of vector called "access vector" [1]. An alarm is issued whenever the similarity between the access

vector and the typical- terrorist- behavior is above some predefined threshold T. The module for this is shown in Figure 2. This inequality is represented as:

$$\max\left(\frac{\sum_{i=1}^{m} (tC_{i1}.t\,A_i)}{\sqrt{\sum_{i=1}^{m} tC_{i1}^2 . \sum_{i=1}^{m} tAi^2}}, \ldots, \frac{\sum_{i=1}^{m} (tC_{in}.t\,A_i)}{\sqrt{\sum_{i=1}^{m} tC_{in}^2 . \sum_{i=1}^{m} tA_i^2}} \right) > T$$

where, C_i is the i^{th} centroid vector, A_i is the access vector, tC_i is the i^{th} term in vector C_i, tA_i is the i^{th} term in the vector A_i, and m is the number of unique terms in each vector. The optimal value of T depends on the preferences defined by the users.

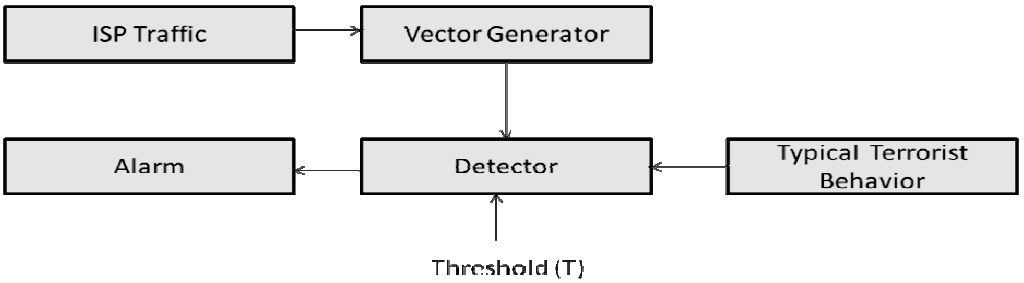

Figure 2: Detecting Terrorists-Monitoring Module

3. DESTABILIZING TERRORIST NETWORKS

Nasrullah Memon et.al developed a hierarchy of a terrorist network to destabilize the network [7]. They converted an undirected graph into a directed graph by an algorithm which uses some centrality measure from SNA literature i.e. degree and eigenvector centrality measure and then converted the directed graph into a tree by applying another algorithm which uses a newly introduced measure called as dependence centrality.
The dependence centrality (DC) of a node means how much that node is dependent on any other node in the network. Mathematically it can be written as:

$$DC_{mn} = \sum_{m \neq p,\ p \in G} \frac{d_{mn}}{N_p} + \lambda$$

Where ***m*** is the root node which depends on n by $\boldsymbol{DC_{mn}}$ centrality and $\boldsymbol{N_p}$ actually is the number of geodesic paths coming from m to p through n i.e. it represents the number of alternative paths available to node m to communicate to p, and $\boldsymbol{d_{mn}}$ is geodesic distance from ***m*** to ***n.*** λ is taken 1 if graph is connected and 0 in case it is disconnected. Hence, it can be said that DC depicts the usefulness of node n to node m for communicating with other nodes of the network.

A node having low dependence centrality might be a key player i.e. leader/gatekeeper as they have direct links to other nodes of the network and they do not need any other node to

communicate with those nodes. The values of dependence centrality can be organized in form of matrix where each row corresponds to a particular node and its DC to other nodes is represented in different columns of same row. When all the values of a particular row are summed, the sum value shows how much that particular node is dependent on other nodes. The lower the sum, the less the node will be *dependent* on other nodes or that node is said to be free-living (*independent)* node. In the same way when all the values of a column are summed, the sum value shows how much other nodes are dependent on a particular node that is associated with that column.

Now, the hierarchy developed using this dependence centrality helps law enforcement agencies to easily identify the leaders and peripheries in the network in order to destabilize the terrorist network. Dependence centrality measure may also be very useful in destabilizing terrorist networks because it shows the nodes which are totally depending on particular nodes. If the nodes are completely depending on the other nodes, they will be isolated (cut-off from the network completely) by capturing the node on which those nodes are depending [8].

5. FUTURE OPPORTUNITIES

Future work in this area can be done in order to destabilize the network more effectively by utilizing more than one centrality measures. Another possibility for the enhancement includes the efficient implementation of Fuzzy Logic and Genetic Algorithms for the classification and identification of terrorist groups within the network. It has been analyzed that the use of Fuzzy Logic makes the detection process synonym to real world, defining the candidate set on the basis of uncertainty in support and confidence framework [3] while the Genetic Algorithm optimizes the detection process making the intrusion detection effective and optimizing the membership function. Further, these membership functions and patterns are stored for future use [10].

6. CONCLUSION

Present paper not only studies about what the terrorist mining is? But also explores the major trends to detect the terrorist networks. The motive of the detection process is to efficiently detect and restrict the inhuman activities. SNA on one hand introduces the detection process through network analysis in the form of a graph and cluster analysis for subgroup detection; while content-based on the other hand talks about the detection using IDS. After the detection of the key player nodes, the nodes are to be destructed from the network. This is done by using destabilizing of the network using destabilizing techniques. The goal of destabilizing techniques is to capture or kill a critical node from a network. This makes the network weak, less adaptive and the network no longer remains as a single entity that is it breaks into components. So some techniques should be established which makes the network adaptive after killing of a key node and connect all the disconnected component in negligible time. Destabilizing techniques should be such that it should be able to capture/kill the key player i.e. the leader and the potential leader in the network so that maximum damage is caused to the network and then the terrorist are unable to fulfill their inhuman goals.

REFERENCES

[1] R.P. Lippmann, D.J. Fried, Graf. J.W. Haines, K.R. Kendall, D McClung. D. Weber (2000)," *Evaluating Intrusion Detection Systems: the 1998 DARPA Off-Line Intrusion Detection Evaluation*", Proceedings of the 2000 DARPA Information Survivability Conference and Exposition, Vol. 2.

[2] Arun K. Pujari (2001), *Data Mining Techniques*, Universities Press.

[3] *German* Florez, Susan M. Bridges, and Rayford B. Vaughn (2002) "*An Improved Algorithm for Fuzzy Data Mining for Intrusion Detection*", Proceedings of NAFIS2002 Annual Meeting of the North America, 457-462.

[4] Jiawei Han & Micheline Kamber (2006) *Data Mining;Concepts and Techniques*, Second Edition, Morgan Kaufmann Publishers.

[5] Pang-Ning Tan, Michael Steinbach and Vipin Kumar (2006), *Introduction to Data Mining*, Addison-Wesley Companion Book Site

[6] Muhammad Akram Shaikh, Wang Jiaxin, (2006) "*Investigative Data Mining: Identifying Key Nodes inTerrorist Networks*".

[7] Nasrullah Memon, Henrik Legind Larsen, (2006) " *Practical Approaches for Analysis, Visualization and Destabilizing Terrorist Networks*", Proceedings of the First International Conference on Availability, Reliability and Security (ARES'06).

[8] Philip Vos Fellman, Roxana Wright(2007) "*Modeling Terrorist Networks - Complex Systems at the Mid-Range*".

[9] Hosseinpour, M.J.; Omidvar, M.N., (2009) "*Detecting Terror Related Activities on the Web with Using Data Mining Techniques*", Proceedings of the Second International Conference on Computer and Electrical Engineering, 2009(ICCEE '09), Vol 2, 152-157.

[10] Shingo Mabu, Member, Ci Chen, Nannan Lu, Kaoru Shimada, and Kotaro Hirasawa,(2011) *"An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming",* Proceedings of the IEEE Transactions on Systems, Man, and Cybernetics—Part c: Applications and Reviews, Vol. 41, No. 1, 132-139.