Covert Network Analysis to Detect Key Players using Correlation and Social Network Analysis

By

Ejaz Farooq

NUST201261299MCEME35412F

MS-12(CSE)

Submitted to Department of Computer Engineering

In fulfilment of the requirements for the degree of

Masters of Science

In

Computer Software Engineering

Thesis Supervisor

Dr. Shoab Ahmed Khan

College of Electrical and Mechanical Engineering

National University of Science and Technology

July 2016

This page is intentionally left blank

Covert Network Analysis to Detect Key Players using Correlation and Social Network Analysis

Author:

Ejaz Farooq

NUST201261299MCEME35412F

Submitted to the Department of Computer Engineering in fulfillment of the requirements for the degree of
MASTER OF SCIENCE IN SOFTWARE ENGINEERING

Thesis Supervisor:

Dr. Shoab Ahmed Khan

Thesis Supervisor's Signature:____

DEPARTMENT OF COMPUTER ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD

JULY, 2017

DECLARATION

I hereby certify that this research work titled "*Covert Network Analysis to Detect Key Players using Correlation and Social Network Analysis"* is my own work under the supervision of Dr. Shoab Ahmed Khan. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred and there is no plagiarized data contained in this research.

Signature of Student

Ejaz Farooq

NUST201261299MCEME35412F

LANGUAGE CORRECTNESS CERTIFICATE

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student

Ejaz Farooq

NUST201261299MCEME35412F

Signature of Supervisor

COPYRIGHT STATEMENT

ACKNOWLEDGEMENT

All praises are for Allah the Almighty because nothing would have happened without his uncountable blessings and love.

I would like to express my sincere and deep gratitude to my supervisor, Professor **Dr. Shoab Ahmed Khan**, for his guidance, constructive and valuable suggestions and comments and kind support throughout my work. His constant encouraging attitude has always been a moral support for me and for that I will forever be grateful.

I owe a lot of my gratitude to my co-supervisor Dr. Wasi Haider Butt and all my teachers. Their guidance supported me during the time study and during research.

I feel particularly indebted to my parents, brothers and sisters for their ever-lasting love support understanding and encouragement especially through the hard times. It was their unshakable faith in me that has always helped me to proceed further. I cannot express my love and gratitude I have for them. Thank you for encouraging me and supporting me in every step so that I can fulfil my dreams.

Finally, I wish to extend a warm thanks to friends and to everybody involved directly or indirectly with my work.

ABSTRACT

A sudden escalation of terrorist events entices many researchers to divert their attentions to counter-terrorism field, and contribute in developing new techniques and methods for analysis, identification, exposing and prediction of terrorist events by using latest technologies. To be more effective, terrorist keep their identity secret in hierarchy of any organization to escape from the eyes of law and enforcement agencies. But to accomplish the fatal activities of terrorism successfully, they must need to communicate with others in network to make plans. The stealthy hierarchy of such terrorist organizations could be exposed by the pattern of collaborations and communication among members. The social network could be defined as "A social collection made *up of social actors like persons or organizations and a compound set of links between these actors". This definition inspires us to treat terrorist networks as the normal social networks, so that different social network analysis techniques could be applied on such networks to extract their hierarchy and useful information, which could help us to predict something about terrorist and their activities. The prediction about stealthy hierarchy of terrorist networks would expose the significance of every node in the network. This could help us to predict something useful in order to destabilize the network, which eventually results in immobilization of terrorist to accomplish their evil intents. From our analysis, we come to know that the traditional measures for social network analysis are not capable or related to above mentioned problem, except from the one the relative degree. The inability of traditional measures is caused by the stealthy hierarchy and also hidden intents of the terrorist. A node in terrorist network might not be prominent but that particular node might be the leader of the network. Removing that node from the network will help in easily destabilizing the network. Current methods of social network analysis mainly focus on the different degree measures considering nodes and edges of the network, while there is still a lot of work need to be done keeping the focus on the communication between nodes. These aspects of social networks like secrecy and lack of analysis to detect importance of node in network using nodes communication inspires us to propose a framework to detect key players in covert network using text mining. This thesis, describes to build a model to find the correlation between a data dictionary and communication of nodes, to evaluate and detect the key players having the highest similarity with data dictionary, which consists of words or terms mostly used by terrorist for their organizational structure, terrorist activities, planning etc. Our research work mainly presents this* *novel model and analysis to detect key player using this model. In this thesis, we used Enron email dataset to test and validate our novel proposed model. In this thesis, we also develop a data dictionary from the selected dataset, as a pre-defined data dictionary was not available. We also devise a preprocessing module, which could be used in any text many application. Till date according to our understanding no such model exist, which is dealing with the conversation data of nodes for key player detection.*

***Keywords** – key player, preprocessing, similarity measures*

Table of Contents

Chapter 1

Introduction

Chapter 1

1.1 Introduction

“Covert network analysis for key player detection using text mining” is a research work to find the key players in a social network of terrorist based on their communication, which is totally a new concept in social network analysis (SNA). In this thesis we will present how to detect key players using communication by finding the correlation between a reference data dictionary which consist of words and terms related to terrorist and communication of nodes in the proposing a framework. So, this is our main objective of this research. Data dictionary should be pre-defined and node communication should be structured are pre-requisites to detect the key players in a terrorist network, and will be used for finding the matching similarities between pre-defined dictionary and communication of any node in network.

Key player detection means that find the influential members of social network, those who play most important role on key decisions of terrorist events or management of the terrorist organization.

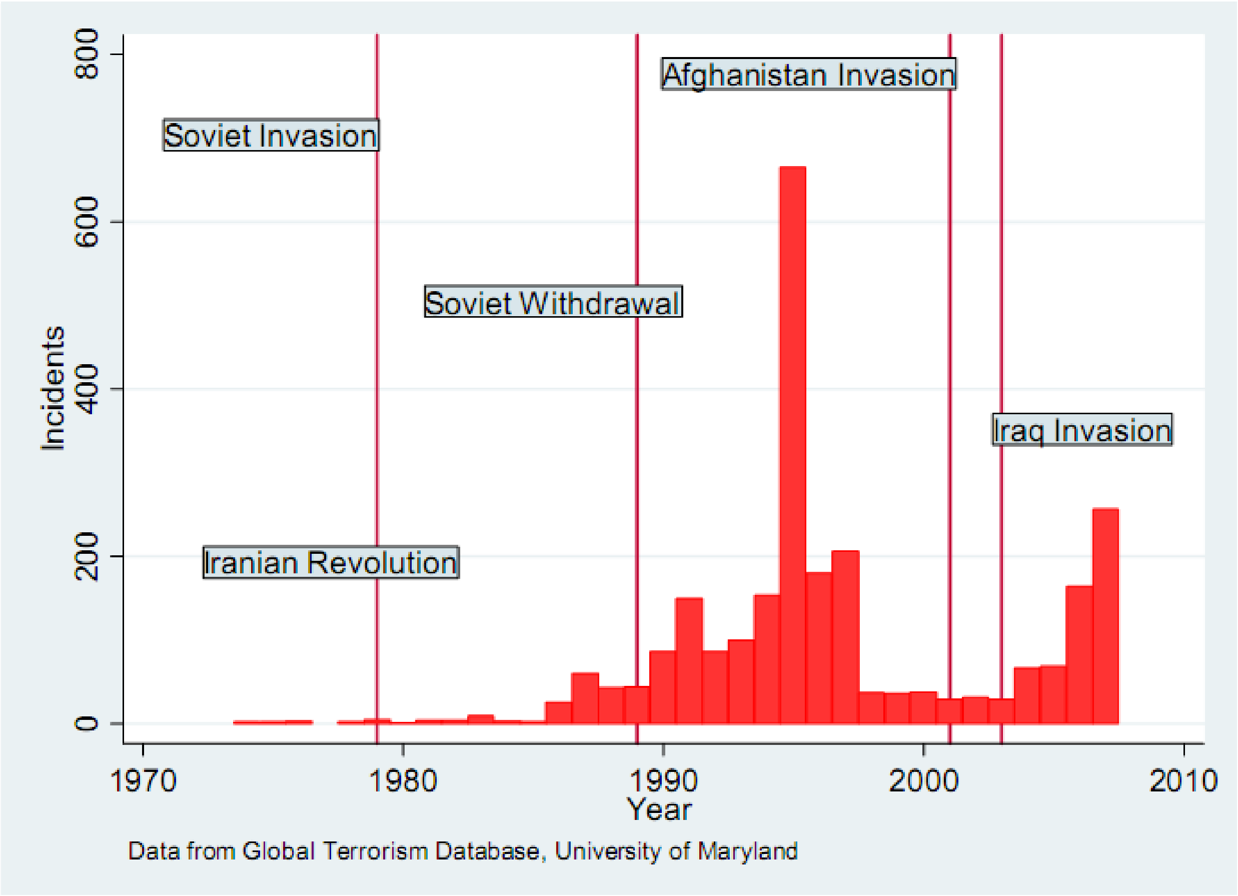
National Security has gained vital importance since the 9/11 famous terrorist attack. The anti-terrorist field demanded specially researchers from Information Technology to contribute in this field to provide support to law and enforcement agencies to fight against terrorist and destabilize their networks. The need for the research in this area has increased and it has many reasons. There were increasing number of terrorist events not only in worldwide but also at national level as well. And the most important reason for Information Technology researchers to work in this field is that terrorist and their organizations are now extensively using the latest technologies, devices and applications for the communication and planning as well. Without these it would not be possible for the terrorist to communicate with others regularly and effectively in this era.

In this thesis, a model is proposed to find out the correlation or similarities between a pre-defined dictionary and communication of any node in the network, so that sone can detect the key player (or a terrorist) in network and predict the future events to some extent based on the dictionary.

1.2 Motivation

As our country is geographically situated in a region where every terrorist act in its any neighboring country will have a great effect on its National Security. Unluckily we are the most badly affected nation from terrorism not only in the region but also in the world as well.

This has created huge potential for research in anti-terrorism field for National Security. So, this was the reason to select this field for research work. Figure 1.1 shows that Pakistan has always been affected by terrorism with problems in the neighboring countries.



Data from Global Terrorism Database, University of Maryland

Figure 1.1 Number of terrorist incidents since 1973 to 2007 in Pakistan

1.3 Terror and Terrorism

It is really very difficult to describe the word “terrorism” in modern era; there are multiple definitions available from multiple sources. “Systematic use of violence and intimidation to achieve some goal” is one of the generally used definition for terrorism. The resolution passed by the UN General Assembly also provided a definition for terrorism which is:

“Criminal acts intended or calculated to provoke a state of terror in the general public, a group of persons or particular persons for political purposes are in any circumstance unjustifiable, whatever the considerations of a political, philosophical, ideological, racial, ethnic, religious or any other nature that may be invoked to justify them”.

Almost all the areas of social life are being affected by terrorism of any society, which includes education, sports, politics, economy and event managements etc. Any kind of terrorism is dangerous for any nation. So far, terrorism has affected almost the all globe. There are a number of terrorist organizations, and terrorist work for them and they have killed all kind of innocent people irrespective of gender and age, no respite for woman, children and old as well. This has caused the destruction of so many societies, which results in helpless people deprived from their homes as well. Some of the worst killing events in the world are listed below which results in casualties of 100 or more people: [1]

- April 16, 1925; Bulgaria, Cathedral bombing in Sophia: fatalities were 160
- May 18, 1973; Siberia, Mid-air bombing of Aeroflot airliner: fatalities were 100
- Aug 20, 1978; Iran, Arson of theater in Abadan: fatalities were 477
- Nov 20 – Dec 5, 1979; Saudi Arabia, Hostage taking at Grand Mosque in Mecca: fatalities were 240- including 87 terrorist killed
- Sep 23, 1983; Over UAE, Mid-air bombing resulted in crash of Gulf Air flight: fatalities were 112
- Oct 23, 1983; Beirut, Lebanon, Truck bombings of U.S. Marine and French barracks: fatalities were 301
- May 14, 1985; Sri Lanka, Armed attack on crowds in Anuradhapura: fatalities were 150

- Jun 23, 1985; Canada, Mid-air bombing of Air India flight off Ireland, and attempted bombing of second flight: fatalities were 331
- Jan 20, 2012; Nigeria, Multiple bombings in Kano: fatalities were 178
- Mar 12, 1993; India, 15 bombings in Bombay: fatalities were 317
- Dec 21, 1988; Scotland, mid-air bombing of Pan Am flight over Lockerbie: fatalities were 270
- Aug 29, 1997; Algeria, attacks at Sidi Moussa and Hais Rais: fatalities were 238
- Oct 28, 2009: Pakistan, bombing at marketplace: fatalities were 119

Table 1.1 shows the death toll caused by bloody incidents of terror in Pakistan yearly.

Year	**Civilians**	**Security Force Personnel**
2003	140	24
2004	435	184
2005	430	81
2006	608	325
2007	1522	597
2008	2155	654
2009	2324	991
2010	1796	469
2011	2738	765
2012	3007	732
2013	3001	676
2014	1781	533
2015	940	339
2016	308	151
Total*	**21185**	**6521**

Table 1.1 Fatalities in Pakistan caused by terrorism 2003-2016

Terrorist who carry out such brutal incidents are understood to work in groups also known as terrorist networks. These terrorist network are affiliated with some extremist organization. As any other normal organization has its interests and goals, such terrorist outfits also work for their benefits and goals, keeping them hidden. These outfits work globally. They do not restrict them to a certain region or area.

US state department has included some highly reputed extremist outfits in the list of terrorist organizations. These includes: "Abu Nidal Organization" (ANO), "Basque Fatherland and Liberty" (ETA), "Liberation Tigers of Tamil Eelam" (LTTE), "Al-Qaida in the Islamic Maghreb" (AQIM) and so on [2]. This list also includes organizations like Taliban, ISIS etc.

1.4 Social Networks:

A collection comprises of social actors for example people, institute or organization and a complete set of relationships among them, is defined as the social network. Social network view helps us to have better analysis of the structure for social bodies [3]. "Social Network Analysis is a Mathematical method for connecting the dots". SNA often permits us to do mapping and measurement of complicated and maybe hidden hominid groups and organizations [4]. In order to discover the numerous exciting characteristics of diverse nature of social networks, social network analysis (SNA) has been applied on multiple applications.

1.4.1 Terrorist Networks as Social Networks

A model used to understand the terrorism is called terrorist network. There are many people involved in such networks and among those people terrorist also involved. [1302.1727] Terrorist network consists of terrorists. To achieve their aim of doing the terror activities, these terrorist used to work with in organize and a well-defined structure. There are two significant issues need to be discussed here regarding this topic. First of all, "why terrorist networks should be counted in social networks at all" and the another is, "is a terrorist network a pure social network?". These question arises because we need to work with SNA on such networks, can SNA measures be applied on such networks or they have different features or specific features apart from the social network.

Why terrorist network should be in the class of the social network, is first discussed here. After the occurrence of devastating 9-11 event, many law and enforcement and the significant agencies shifted their attentions towards these terrorist groups in great deal, bigger than ever before and also the attentions of the researchers working in the anti-terrorism field to cope with them, there was also a reason that number of terrorist attacks increased significantly after that. After 9-11 event names of terrorist organizations become famous among common people, because their names appeared in media more frequently. By performing the analysis of the available information from all sources, this could be determined that terrorist outfits having some their own precise features and aims works as the classis organization. We can map terrorist groups to social network as the terrorist belonging to any organization can easily be mapped to nodes of social network and the collaborations between them could be considered as the links/edges of the social network. These collaborations could be anything like fellowship, relationship, communication and having training together etc. It is obvious that terrorist have to be very careful while collaborating with others and also using different resources in order to accomplish their hidden aims. Now discussing the second issue of considering a terrorist network a pure classic social network. But from the studies and research in this area, we find out that it is not the right case. These are two different networks. Although terrorist networks have their own particular features different from the classis social network, but still they can be considered as social network. There is a notable difference often can be seen in the behavior of the social network and the hidden networks [5]. In covert networks, mostly different important links remain in passive mode which help them to remain hidden and only come to active mode when they required, as this is not the case in social network [6]. By considering and understanding of the current centrality measures, most active nodes shall be considered as the key actors in normal social network, but in hidden networks due to privacy this is not the right case.

1.5 The Key Player Problem

The detection of the key player is one of the most important aspect in social network analysis (SNA). The most important node of the social network is considered as the key player of the network. The nature of the network could determine the importance criteria of nodes. Individual nodes are easy to find in social network who has no interest in hiding their importance in the network, as compared to those individual nodes who tend to keep their secrecy in the network.

They do not want to show their identities, their roles and their relative importance in the network for the sack of the network stability. So, terrorist networks lies in such category. In terrorist networks, to accomplish a secret goal different individuals connected together secretly. Members of terrorist network always remain very careful and watchful from being trapped. So, the problem is to detect the key actor from the network whose exclusion can disable the whole group from accomplishing the secret objectives.

1.5.1 Centrality

To do the research in social networks, theory of centrality has very important role to understand the organizational and team conduct in social networks. Key central figures in the network has full control over the key decision making and on the flow of information. "But, it is not yet clear the linking between the real world phenomenon of centrality and mathematical measures of centrality" [7]. The highest values of centrality of nodes signifies that those actors has the maximum hierarchical significance in relative network, and are considered to be playing and important role in real-world activities and in cyber activities as well. Many different nature of networks work under this rule [3].

1.5.2 Communication of terrorists

In this era terrorist organizations and individuals are using different means of information technology to collaborate with each other and within organization. It is certain that to keep themselves hidden from law enforcement agencies who keeps eye on social media to detect terrorist, they surely communicate in code-words mostly. To fool authorities, Al-Qaeda is famous for using code-words as well as religious terms in their communication. This is also stated as covering speech. Language security is a frequently used policy. For example, writing emails in such code-words that would not be easy to translate for authorities same is the case with the dialogs or website literature.

Where centrality measures are unclear, communication can be a great source to detect key players in terrorist networks.

1.6 Problem Statement

Social network analysis, defined a data dictionary for terrorist communication, developing words vector of terrorist communication, proposing a novel model to detect key player in covert/terrorist network using text mining based on nodes communication data and data dictionary.

1.7 Thesis Contributions

In this thesis following contributions are made:

- Development of data dictionary based on the words or code-words used by terrorist for communication
- A vector creation model of terrorist's communication to find the correlation or matching similarity with the pre-defined dictionary also called pre-processing
- A novel model to find the similarity between the pre-defined dictionary and terrorist communication data for classifying a node a key player or not has been proposed
- Pre-preparing a training dataset for experimentation purpose

1.8 Thesis Organization

The thesis report is consists of the 5 chapters. **Chapter 1** has been discussed above, which outlines introduction and problem in detail. While **chapter 2** discussed the related work done in the anti-terrorism field. It outlines what is social network, analysis of social network, current standard measure of social network, how those measures are used to detect key players in counter terrorism field, what others work have been done in this field for detecting are discussed in this chapter. **Chapter 3** outlines the proposed framework and all the contribution made in the thesis. It describes implementation of the framework. **Chapter 4** discussed the results, analysis and method for detecting key players from the model discussed in chapter 3. **Chapter 5** discussed the work done and also the future enhancements to carry this work to ahead, it also outline the contributions of the thesis in brief form as well. In the end references was also mentioned.

Chapter 2

Literature Review

Chapter 2

In this chapter we will be discussing the study done related to the field of social network analysis along with the basic concepts involved. Apart from a brief overview on afore mentioned study we will discuss about the key player detection as it is considerably the central point of this research. We will discuss in detail about figuring out and analyzing the key players in the terrorist network specifically.

2.1 Social Networks

Before going into deeper details, it's important to review what a social network actually is. It might be defined as a social group that consists of human beings and/or organizations, which can be said to be the social actors of a social network who are related in some way or other. The relationships between these social actors is the key in a social network.

It can be better explained with a graphical representation which may illustrate the relationships between the social actors involved in a certain social network. Every actor is represented as nodes or vertex while the relationship between them is illustrated by connecting the actors with an edge. A graphical notation helps the social network analysts to better understand the network and the relationship between its actors. The visuals are easier and a quick way to grasp the bigger picture of any network. But sometimes to get to have a closer look the analysts might need to take a mathematical approach which basically involves the theoretic concepts of the already discussed graphical view of a network.

Social networks can turn out to be very useful in identifying and calculating correlation among multiple entities or social networks and also in symbolizing the associations between these entities. [8] Social networks have been effectively applied in various fields like sociology, biology etc. [9]

A very good example of the application of social networks is the association between people working together, which is shown in the following matrix. The headers represent the actors (nodes in graphical illustration) while the relationship between these actors is shown in details. If a node is not related to another it is shown by 0 while if there does exist a relationship then the cell shows a 1. The diagonal values are null because a node cannot have a relationship with itself.

	Shiraz	Saeed	Umer	Imran	Mohsin	Qasim	Irfan	Asim	Salman	Ibrar	Rashid	Razi	Ejaz
Shiraz	-	1	1	1	1	1	1	0	0	0	0	0	1
Saeed	1	-	0	0	1	0	1	0	0	0	0	0	1
Umer	1	1	-	1	1	1	1	0	0	0	0	0	1
Imran	0	0	1	-	1	0	1	0	0	0	0	0	0
Mohsin	0	0	1	1	-	1	1	1	0	0	0	0	1
Qasim	1	0	0	0	1	-	0	1	1	1	1	0	0
Irfan	1	1	0	0	1	0	-	1	0	0	1	0	1
Asim	1	0	0	0	1	1	1	-	0	0	1	0	1
Salman	0	0	0	0	0	1	1	0	-	0	1	0	1
Ibrar	0	0	0	0	0	1	0	0	1	-	1	0	0
Rashid	0	0	0	1	1	1	1	0	0	0	-	0	0
Razi	0	0	0	0	0	1	0	0	0	0	1	-	0
Ejaz	1	1	1	0	1	0	1	0	1	0	0	0	-

Table 2.1 The social network data showing colleagues relation

Figure 2.1 shows the graphical representation of the above mentioned matrix of relationship between colleagues regarding work.

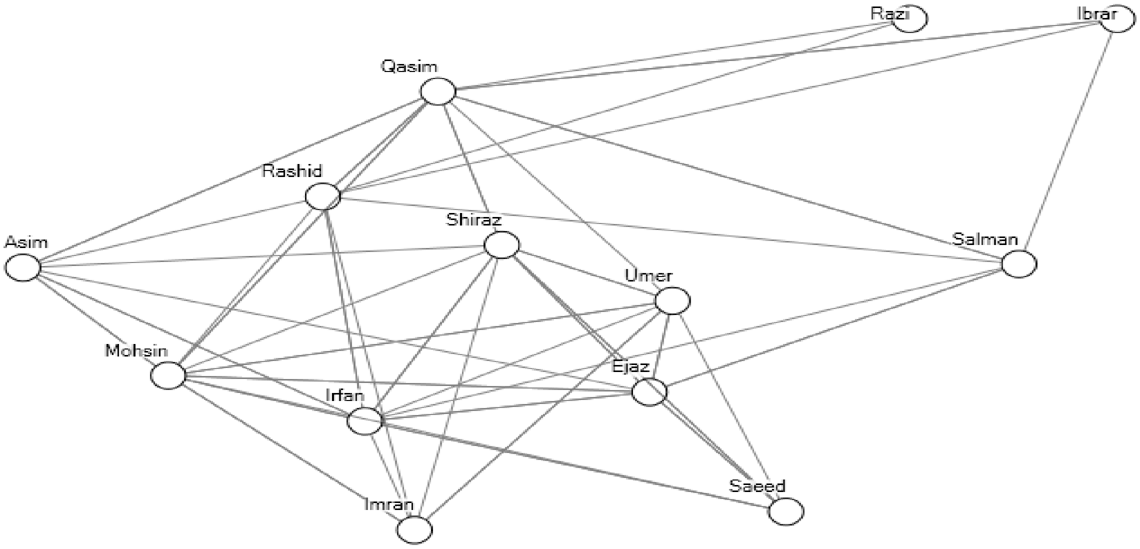


Figure 2.1 Network of colleagues working relationship

Chapter 2

2.2 Social Network Analysis (SNA)

Social Network is defined uniquely with different perspectives by many researchers. It basically includes extraction and manipulation of existing data in any give social network. Krebs describes SNA from the perspective of relationship between people as how these relationships are mapped and measured [8]. Similarly, Scott also defines it from the relationship point of view but specific to collection of procedures to explore relational features. Freeman however thinks of it as effective techniques to discover hidden patterns from the data obtained from the interaction between the people of social network [11].

When it comes to SNA, it does not focuses on the nodes but instead on the relationships between all the nodes. However, the attributes of the nodes sometimes becomes essential to help in apprehending the social behaviors and analysis of specific social phenomenon. This why it is important to always include the attributes of the nodes in the study even though SNA is primarily focused to interaction between the nodes and the structural features of these nodes. The SNA analysts mostly motivated to find important features like composing subgroups from a network based on the relational aspects or how the network connects etc.

Thus, the tasks for SNA are typically divided into 3 broad categories. The first category includes the actors who work dedicatedly to one main task for example people working in a sales department or a group of individuals focused to one goal or possessing common attributes. The second category includes the relation between the individuals based on how they are connected. This second category is the main focus of SNA. There always has to be a reason for which the analysis is being conducted and this reason is what determines the nature of relationships between the actors. Relationships can be concluded from the communication between the actors which also means the links or ties between individual actors. The third category are the properties or attributes of the actors and the relationships identified. The attributes are the factors that might affect an actor or a relationship between actors. This implies that only those attributes are chosen that affect the links between the actors in some way or other. Other attributes are simply ignored by the SNA.

SNA is not easy to perform manually and might be very hectic and time taking but after much innovation in the information technology the researchers have managed to develop easy to use state-of-the-art software systems to make SNA convenient, efficient and less error prone. In addition to proprietary tools there are open-source tools available to conduct such researches. Using these tools researchers can carry on the analysis with ease by applying several built in SNA operations on all types of network circles. Moreover, the researchers are able to visualize networks under study with graphical representation.

There are all types of tools that can be used with a variety of perspectives and meet the criteria of different practices. NetDraw, NetMiner and Pajek have capabilities to serve good for the researchers who believe that human eye is crucial for best analysis. These tools illustrate the social networks visually in an appropriate and useful way. Other tools include features like text reports that are generated based on the SNA measurements, attributes and inputs. UCINET, MultiNet and Agna are good reporting tools. The best perspective however is to provide testual and visual illustration to accommodate all types of studies. This hybrid approach can be very useful in concluding accurate results. There are tools that help with statistical analysis as well for example StOCNET. But all these approaches and the tools so developed still need improvement and that can be achieved by further extension research and development on these tools. Such efforts are still in progress by various scientists. When analyzing small networks these tools give better results but the problem arises with larger networks. Large networks are complex and dense and hence, require efficient tools that are built on huge knowledgebase. The exiting tools have different limitations in different perspectives when the number of nodes and relations that can be analyzed extend.

2.3 Application of Social Network Analysis

As discussed in the above sections, a structure that defines individuals or organizations and the connections between these actors is a social network. Email traffic, criminal networks, social media networks, sales and distributors' network, all can be modeled as social networks. Therefore, analyzing any of such networks will be called social network analysis or SNA. As discussed earlier, SNA measures and maps the interactions between the actors of any social network. The figure below illustrates the network of AIDS.

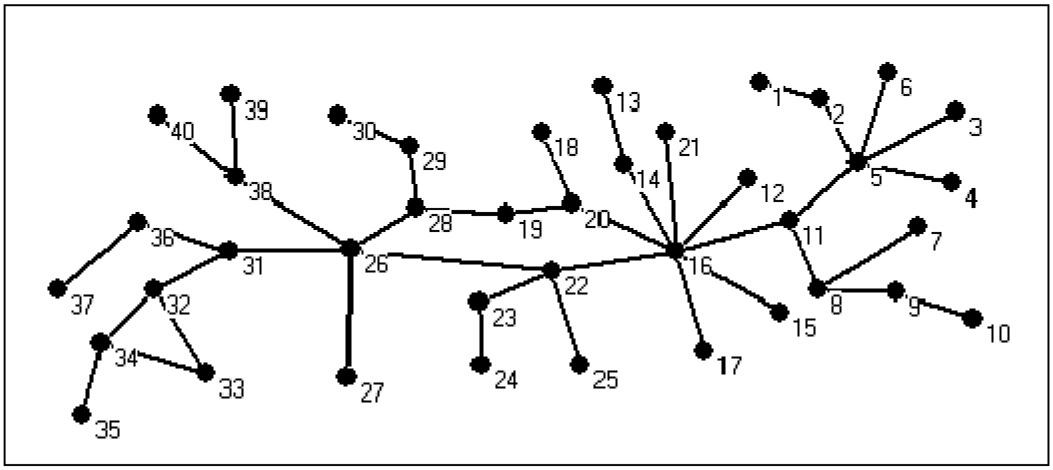


Figure 2.2 Network representing the relations in AIDS network [32]

With SNA, we can analyze the relationships of a social network mathematically and visually. When the SNA is conducted over the relations and network of an organization, it is called ‘organization network analysis’ (ONA). Studying the network of an organization is very complex and so is further extended to the study of sub-groups.

SNA gives answers to studies like the conflicts that can arise in a network and what impact it has on the relationships between the actors, how quickly any piece of information flows within a network and what path it follows, what is the scale of overlapping cliques in a network etc. A lot of features together make a dataset for any social network. Social networks data contains information including the actors and social ties between actors which together constitute a social relational system [3]. Having said that, it is very important to detect the most important actors and their relations with other actors in the whole network. This way the unique characteristics of the social network can also be concluded. This type of study can be very helpful in many areas. E.g detecting the most influential actors in a terrorist network can help in control the way the communication of the network flows and also help in disrupting the whole network. Detecting the most influential distributors in a sales and distribution network can help in obtaining more revenue by using the important distributors. This concept in SNA is represented by the term Key Player which will be discussed in the next section.

2.4 Key Player Detection

Key Player Detection plays a crucial part in SNA of a social network. Depending on the types of social networks a key player can have different roles and characteristics. It might have multiple responsibilities in that network which define its significance. This implies that the key player can be defined differently depending on its roles and responsibilities in a particular social network. Identifying these responsibilities is crucial as these responsibilities and characteristics are what define our social network as a whole. It is as critical to define the criteria for every social network on the base of which its actors are evaluated as key players or not, depending on the type of analysis. This criteria also gives us a lot on insight into the network itself or vice versa.

Stephen P. Borgatti followed the approach proposed by Friedkin [13] and supports his approach that the detection of key player involves solving two sub problems. First, to identify the subset of actors which if eliminated from network causes abruption, second, the subset of actors that are connected to maximum actors in the network. These researches inspire us to follow the same approach to conduct this study. Borgatti, extended the approach and introduced two terminologies based on the two problems, Negative Key Player Problem (KPP-Neg) and Positive Key Player Problem (KPP-Pos) [14].

KPP-Neg represents the amount of integrity in the network with the presence of the subset of key players in the network and the reduction in that amount when this set of actors is eliminated from the network. This from another perspective shows us how significantly the key players play the role of keeping the network actors together hence maintaining the stability of the network. Borgatti defined KPP-Neg as “the degree of dependence of the cohesiveness of the network on the presence of its key players”. The significance of the key players is measured in terms of the disruption caused, i-e the amount of cohesiveness of the network after the elimination of the set of key players. This assures the discoordination among the network for example among terrorist groups. KPP-Neg can have major impacts on the strength of a network.

KPP-Pos represents the amount of connectivity between the key players and the ordinary actors in the network. This can mainly be used in problems where there is a need to spread information, behaviors, trends etc. It may also be used to stop something from spreading among the network by identifying the key players and eliminating them, for example, disease. The KPP-Pos can have an impact on the flow between the network actors. Borgatti defined KPP-Pos as the degree to which the key players are embedded in a social network.

Generally defining the two approaches, Borgatti uses *k* to represent the subset of actors in both cases. In KPP-Neg, *k* is the subset of actors which can cause the network communication among the other actors maximally disrupted when removed. In case of KPP-Pos, *k* refers to the subset which is extremely connected to the rest of the actors [14].

Two detect the subset of key players, centrality serves as an important factor in SNA. The centrality can be referred to as the importance of each node in a network. First the importance is identified and each individual is scored on a scale. Then a threshold is defined and any individual exceeding that threshold is selected to be a potential key player. But this procedure is not as simple as it seems and comes up with its own problems that need to be solved.

2.5 The Problem of Centrality

In 1948 Bavelas, had introduced the theory of centrality as applied on human collaboration. The communication takes place in small group was his major worry. He presented a hypothesis, which discussed the relationship between structural centrality and effect in group processes [15]. In late 40s, Bavelas directed the first research application for centrality at MIT. The first ever studies were done by Harold Leavitt (1949) and Sidney Smith (1950). Centrality defines the level of individual's center position in the network.

In this section we discussed the centrality measures in a network to find the most important node.

2.5.1 Degree Centrality

It is the simplest centrality measure. DC calculate the importance of a node by considering the direct connection of a node [17]. The formula to calculate the degree centrality with the help of adjacency matrix is given below:

$$\sigma(v) = \sum_{i=1}^{n} \mathrm{a}_{iv} \qquad \mathbf{(2.1)}$$

In above equation, v denotes the index of node for which centrality needs to be calculated and i represents the adjacency matrix index. Degree centrality just count the number of its neighbors.

For undirected network, it simply counts the number of neighbors for each node. But there are two type of degree exist for directed network. The in-degree is one, which calculates the no of incoming edges towards selected node. While the other is out-degree, which calculates the no of outgoing edges from the selected node. A node has highest degree values if it has the highest no of connections. High value of degree centrality represent the central position of a node in network.

2.5.2 Weighted Degree Centrality

The degree centrality was extended for weighted networks and it is now known as the node strength [18]. This centrality measure calculates the centrality by summing all weights of the connected to the selected node, for which centrality is being calculated. Following equation shows how to calculate the weighted degree centrality:

$$\sigma(v) = Strength(v) = \sum_{i=1}^{n} \mathrm{w}_{iv} \qquad (\mathbf{2.2})$$

Where w represents the weighted adjacency matrix. If a node i is connected with node v, then w_{iv} will have value greater than zero. This value represent the weight of the relationship. For binary networks weighted degree and degree centrality will be equal. But there will be difference in weights for weighted networks.

As degree and strength both work on the participation of a node in adjacent network, while doing research on centrality integrating both measures is of high importance [19]. By combining these two, a parameter α is used [19] which concludes the importance of the total number of links as compared to their weights, and this devise a another degree centrality measure.

2.5.3 Closeness Centrality

In degree centrality only directly connected nodes are taken into consideration, this creates a gap for not considering global information. To cater global information, Closeness or Distance centrality was introduced [20]. It checks closeness of a node with other nodes in network. Closeness centrality will have a higher value for those nodes which are at shorter distance to others nodes. Following equation represent how to calculate the distance centrality.

$$\sigma(v) = \frac{1}{\sum_{i=1}^{n} \mathrm{d(v,i)}} \qquad (\mathbf{2.3})$$

Closeness centrality defines how quickly a node can access information in a network through other nodes. A high closeness value shows that a node lies at a shorter path in a network, and can reach to those nodes quickly for information gathering. Such nodes can also have more visibility on the activities of network, can easily control the effectivity of the c.

Shortest path of the network which are also known as the geodesic distance usually used for the calculation of closeness centrality. So, considering the distance, nodes that have shortest path to reach other nodes are tend to have high closeness centrality.

2.5.4 Betweenness Centrality

Between centrality defines the centrality of a node, if it is found to be on many shortest paths between pair of nodes. This centrality measure is based on the idea that interactions between two nodes i and j connected indirectly largely depends on the nodes that located between them. The following equation shows how to c the betweenness centrality.

$$\sigma(v) = \sum_{i=1\ i\neq v}^{n} \sum_{j=1, j<i, j=v}^{n} \mathrm{g_{ij}(v)/g_{ij}} \qquad (\mathbf{2.4})$$

If a node often comes on shortest path between pairs of other nodes, then its betweenness centrality will be at high. Such nodes could be considered as the gateways. Because of such behavior, these nodes also have control over the data flow in network mostly. So, such nodes are tend to be the central figure of the network and could easily destabilize the network if removed. To dislocate the communication in network, nodes having high betweenness centrality can be removed. If working of any network depends on the data flow, then betweenness centrality will be the best measure to find the most important node of such network.

2.5.5 Eigenvector Centrality

Eigenvector centrality discussed that if a node is connected to a more interconnected node, this will have a great impact on its centrality as compared to a node which is connected to a less interconnected nodes. Following equation gives us the EC of a node [23]:

$$\sigma(v) = 1/\lambda_{\max}(A) \sum_{j=1}^{n} a_{jx}.x_j \qquad \textbf{(2.5)}$$

With $x=(x_1,\ldots,x_n)^T$ referring to an eigenvector for the maximum eigen value $\lambda_{\max}(A)$ of the adjacency matrix A.

Eigenvector centrality could also be defined as it is a proportional to the sum of all centralities of node's neighbors. This causes the node to gain high importance either by connected to the nodes that are highly important nodes themselves or by connected to many other nodes.

2.5.6 Valued Centrality

Valued centrality is known to be the alternate of the closeness centrality [24]. For valued network and the strength of links in network, valued centrality was proposed. But later on it works for normal network as well [7]. Its definition make it like closeness, but it different from closeness. Formulation of valued centrality is given below:

$$\sigma(v) = \frac{1}{n-1}\left(\sum_{v\neq i}\frac{1}{d(v,i)}\right) = AVG_{v\neq i}\left(\frac{1}{d(v,i)}\right) \qquad \textbf{(2.6)}$$

By taking the average of the closeness values results in valued centrality.

2.5.7 Jordan Centrality

In 9th century Camille Jordan discovered the centrality (Jordon Center), using this concept Jordon Centrality was proposed by the Hage and Harry. Jordon centrality focuses on the largest distances of every node to find the importance in the network. Following equation gives Jordon centrality:

$$\sigma(v) = \frac{1}{MAX_{i\neq v}\, d(v,i)} \qquad \textbf{(2.7)}$$

2.5.8 Flow Centrality

The alternative of betweenness centrality was proposed as "Flow Centrality" and it is considered appropriate for valued network [26]. As in betweenness centrality location of nodes are of high importance, that a node often comes in between different pairs of nodes. It's a broker type role for a node. Such a node is meant to be a powerful node who has control over the flow of information.

If such a node denies to pass the information through it, the pairs of nodes depending on this nodes for communication must have to stop their communication through that node. Flow centrality discussed that all paths connecting two nodes will be used by that pair of nodes, instead only depends on the shortest path. So flow centrality is calculated by the amount of complete flow between two connected nodes using all paths connecting them that occur on path of which a given node is part.

Then the respective measure adds up for each node depending on the involvement of that node in flows of network between all pairs of nodes, instead for just shortest paths. Based on the magnitude and density of the network, the measurement number will increase for flow centrality.

Following is the equation to calculate the flow centrality:

$$\sigma(v) = \frac{\sum_{i\neq j}\sum_{i<j\neq v} m_{ij}\,(v)}{\sum_{i\neq j}\sum_{i<j\neq v} m_{ij}} \qquad \textbf{(2.8)}$$

2.5.9 Trust centrality

Most centrality measures uses node’s positon in network and then the interaction to find the importance of the node. Trust centrality focuses on trust between nodes, rather than communication of nodes and follow the edges to find the importance of node. It’s a personal kind of centrality. A node need confidence that it’s shared information with other node will be safe and will not be misused.

Based on the privacy setting in the network for each node, trust centrality was proposed [27]. Trust values for each node in the network are calculated and summed. Depending on the privacy settings of all nodes for the selected node, a trust function calculates the trust value. Privacy setting shows which node is trusting which one in the network. So, based on privacy setting trust is computable [27]. Following equation computes the trust value:

$$\sigma(v) = \sum_{u\in V}(w_1\,(u) * \tau(v,u)) \qquad \textbf{(2.9)}$$

If values of trust computed for other nodes is high towards a node, then it will be more trust worth and powerful and if it has low trust value from others then it will be not that important node of the network.

2.5.10 Dependence Centrality

The measure of how much a node is dependent on any other in the network is called the dependence centrality [21]. Following equation gives the dependence centrality.

$$\sigma(v) = \sum_{v \neq u, u \in G}^{n} \frac{m(v,w)}{N_p} + \Omega \qquad \textbf{(2.10)}$$

In above equation v denotes the node having dependency on node u, total shortest paths from v to w through node u is denoted by Np, geodesic distance inverse 1/d(v,w) from v to w is denoted by m(v,w). The graph's connectivity determines the value of Ω, in case of connected graph the value will be one and for disconnected it will be 0.

For experimentation purpose this concept has been applied on the network of 9/11 hijackers [28].

2.6 Other Ways of Finding Important Nodes

Different researchers have proposed some other measures to find the key players from the network rather than by using the traditional centrality measures. Some of the proposed measures are totally new methodologies while some measures still based on the basic SNA measures.

2.6.1 Relative Degree

Relative degree, a totally different measure from the traditional SNA measures. It was designed to intelligently identify the leader of the terrorist network, who supposed to carry some distinct features as compared to other nodes. Group leaders always possess some abilities of high qualities to keep themselves hidden from law and enforcement agencies and able to operate group intelligently. Such leaders are tend to make just plans for all sort of activities and communicate with other responsible to carry out that plans. Each of the SNA measures works on a particular features of node, and devise the most important node considering that feature. For example degree centrality measure only considers number of links between nodes. So, traditional measure would failed in case to find the leaders of terrorist group. Relative degree concludes three hypothesis to find the terrorist group leader generally.

- The group leader will always have less connection because of secrecy
- The leader will definitely be connected to the top degree member for continuity of its plans, monitoring, management of the group

- There will be a very strong relationship between the leader and the highest degree node

Relative degree is defined as "Ratio of degree of maximum degree first hop neighbor to the degree of the node under consideration multiplied with weight of the tie connecting the node under consideration and the maximum degree first hop neighbor". In particularly, keeping in mind the weighted networks this measure was designed. Following equation give us the relative degree.

$$\mathrm{RD(v)} = \frac{\mathrm{Max\big(deg(i)\big)}}{\mathrm{deg(v)}} * \mathrm{Weight\ (Max\big(deg(i)\big)} \qquad (\mathbf{2.11})$$

Where v is the under consideration node, while i represents the set of first hop neighbors.

Below mentioned is the algorithm designed to calculate relative degree.

```
RELATIVEDEGREE (Node N)
   begin
                        max:=0,weight:=1
     foreach (Node m in AdjacentNodes of N)
       if (m != N)
         if (max < Degree(m))
           max := Degree(m)
                                                  weight := weight(m,N)
         End if
       End if
     End foreach
     if (Degree(N) != 0)
        return (max / Degree(N)) * weight
     else
       return 0
     end if
   end
```

Figure 2.3 shows a network based on 9 nodes and 9 relationships as well. From the network, we can conclude that node "C" is connected to 5 members directly, which is maximum in the network, so it has the highest degree. After the application of relative degree on the nodes of network, node "F" is found to be least connected and also connected to "C". Node "F" is detected as the key node. The relative degree for node "F" is calculated as degree of C multiply with degree of F, which results in 5, so relative degree of F is 5/1*1 = 5.

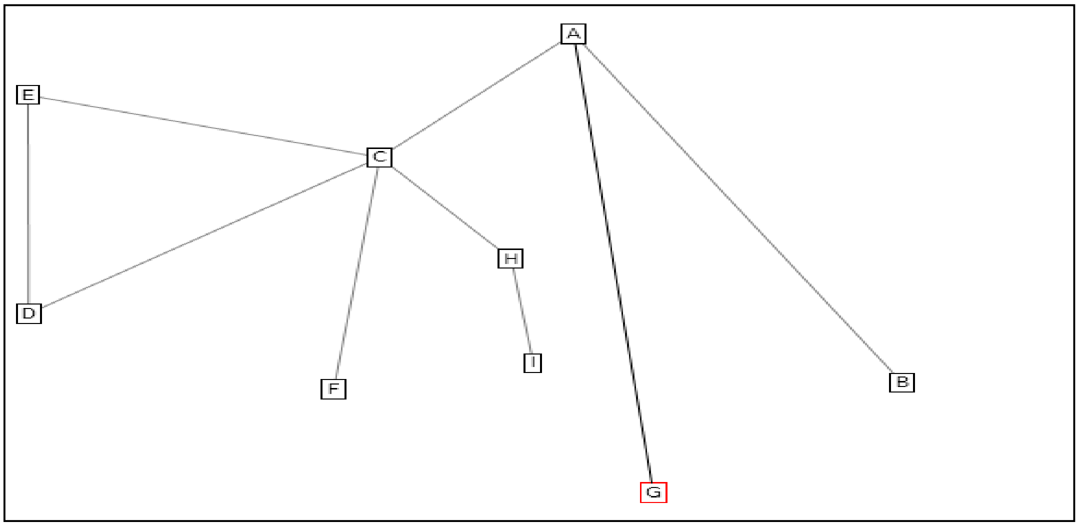


Figure 2.3 A simple network

2.6.2 Using Graph Entropy

Jitesh Shetty and Jafar Adibi discussed the Graph Entropy to detect the key player from a network [30]. The graph entropy they used for their discussion has been taken from the Korner definition [31]. To find the key node from the given network, they discussed and information theoretic model. This model consists of the usage of information theory along with statistical methods of text mining and NLP.

For the evaluation of the model, Enron dataset was used by researchers. In graph of the dataset, nodes represents actors and actions denoted by edges. Actions could be anyone from meeting, email or phone call. The authors categorized the actors into three groups like leader, mediators and followers. Then the comparison of this model with betweenness centrality was taken. This leads them to conclude that when using such approaches, they get poor replies if leaders and followers exist in the network.

2.6.3 Game Theory

For the development of the centrality measures in SNA, game theory was proposed [32]. Game theory discusses the models that uses competition and collaboration. How it will be beneficial for different players while working in teams, such circumstances have been studied in game theory as well. The development of secret networks and this theory has been compared in [33]. Terrorist also work in collaboration to accomplish their evil intents. For completing a successful terrorist event different tasks was required to complete first. These tasks are assigned to different members, and they communicate synchronously to accomplish the target. This leads the secret networks to do the planning and recruitment using communication network [34]. So, it is not important that what the best way to operate for these terrorist is. But the techniques of power allocation among these terrorist is very important. So, for the analysis of the power allocation among the members of the network, guidelines of game theory allocation can be used [35].

With the usage of the cooperative game theory we can model the terrorists involved in an alliance [33]. The combination of the graph theory and the concept of the cooperative game theory, describes a centrality measure which on the basis of the network topology outlines the secret group structure. The amount of money for coalitions gained using cooperating, indicate the value for coalitions in trafficking network for drugs [33]. So, we can detect the key player if we analyze the application of the connectivity games on such networks. To improve the centrality measures for secret networks, authors of [33] have used shapley value concept based on the cooperative game theory solution. Another quality work based on the shapley value has been done in [36].

2.6.4 Use of Behavioral Profiles

A new concept of using the behavioral profile to find the key players of a network is also applied in SNA. Behavioral profiles can be created with the usage of semantic graphs. Semantic structure of nodes and links between them can be denoted as semantic graph. A structure which uses at least one relation for connecting two nodes is called semantic structure. Any individual, location or event are types of the nodes and relation is represented by the link between two nodes. Relationship in sematic graph can exist between different types of nodes e.g. a relation between a location and event. Two nodes can have multiple relation between them due to this nature of semantic graph, these are also known as Multi Relational Networks as well.

For creating profiles an un-supervised framework named SoNMine was introduced by [37]. All the relations of a node are taken into consideration for creating the profile. Semantic graph is used as an input for creating a profile by S. Karthika et al in [37]. The analysis on the behaviors of nodes are carried out after the creation of the semantic profile. Semantic profile is described as a group of shortest paths produced through the method of variable relaxation. Outlier detection was carried out on generated sematic profiles and the highest communicated nodes detected from the outlier are declared as key players.

2.7 SNA in terrorist networks

SNA as discussed before involves studying structures, their attributes and behaviors and thus has been applied to various fields studying social structures. One major and very critical field is studying the terrorist organizations or networks. In this era, when we are facing mass-destruction and major loss of lives due to terrorism there is a dire need to analyze such terrorist groups and take steps to destabilize and eliminate these networks from our today world. SNA plays a crucial role in this aspect, specifically due to its concentration towards social structural analysis. Terrorist networks have appeared to be vast spread unidentifiable structures with hidden objectives. Krebs has done very informative and useful research in this area and has quoted research on illegal and secret networks by social network theorists [6]. Malcolm Sparrow performed extensive research in 1991 and had extensive knowledge of SNA relating to criminal activity. He identified three major problems in analyzing such networks. First is the incompleteness of information regarding unidentified individuals and interactions that are unlikely to be discovered by the analysts. Second, is the identifying a strong boundary as in who may or may not be involved in the network. This information is hard to find. Last but not least, these networks never remain static and keep changing randomly and rapidly. So identifying the nature of the network which also plays a significant role in key player detection becomes very difficult. Keeping in mind these points, Sparrow insists on focusing on the variations of the strength on the identified links over time and tasks is more effective than focusing on the identification of links between individuals. Baker and Faulkner, however, carried on another research in 1993 and proposed to extract relationships from the archived data collected in the past. While, Erickson in 1981 recommends that the members of these networks happen to be connected from past and they are likely to have strong and enduring relationships. The links between the members prior to involvement in the networks is also very important and these relationships are hard to identify.

Krebs carried on his research on the 9-11 terrorist network data and concluded that the undercover network could have been identified if only the links between the schemers were known. The business organizations analysts use four measures to analyze the networks. They focus on trsut between the actors, tasks performed, money & resources involved, and the strategies and goals motivating the networks. Krebs recommends using these measurements might be effective in this type of networks as well.

A lot of researches have been made to contribute in this area. These contributions motivate the application of SNA in counter terrorism. Some of these contributions are discussed in the following sections.

2.7.1 Detection of Chain of Command in Terrorist Cells

When Jonathan D. Farley conducted his research in the area, he came of the idea that to destabilize a network it is very important to identify the hierarchy of the network and the chain of commands. By removing the actors from within a hierarchy it would be easier to stop the flow of such commands to the main soldiers and hence that would stop the major tasks from being performed. He believed that the order followed from the top most hierarchy to the lowest layer and the lowest layer was the one that included suicide bombers, hijackers etc. in a terrorist network. The graphs that were used for the representation of the network ignored this very important aspect of hierarchy in the network which motivated Jonathan to propose use a structured diagrams based on ordered sets that represented all the known layers of a network. This mathematical approach played an important in the destabilizing of terrorist groups [38].

Various studies have been conducted to propose breaking of networks based on their structural properties. But breaking networks does not neutralizes the actors. It only makes them weak but threats still remain. Also, with time these networks regain power by reconnecting and identifying their weak points they can even get stronger. Jonathan's proposal was to identify the hierarchy so as to separate the leaders from the followers. Removing the leadership would certainly cause the network to neutralize.

This approach helps in saving time, energy and resources, both human and monetary, by only identifying n number of actors that are most effective and whose removal will cause the network functioning to stop hence preventing future risks.

2.7.2 Matrix Decomposition

Researches have proved that matrix decomposition technique is rather effective than techniques like clique detection, centrality measurement etc. MD can be extensively used in link and social network analysis. This technique can help in detecting the importance of actors in a group by using relationships data. Three types of MDs can be used according to a study [39]. Singular Value Decomposition, which is an easy way to break or decompose a matrix into simpler and meaningful pieces of information. It eliminates the less important attributes and leaves the important aspects hence reducing dimensions. In the study included in this paper [39], SVD is used to as a tool for graph partitioning and to find irregularities in data. Second type of decomposition that can be used is Semi discrete decomposition, which is widely used in researches as a tool for clustering data after removing noise by using SVD. Third decomposition method is Independent Component Analysis which again converts a multi-directional dataset to a set of subcomponents based on all the directions. It is used to partition similar data on graphs as cliques or sub-groups.

The basic link analysis technique uses the nodes as key factor and focuses on identifying nodes and ranking the nodes before learning about their relationships. Matrix decomposition overcomes this shortcoming of the link analysis technique. D. B. Skillicorn used the matrix decomposition is his research to analyze the famous terrorist group Al Qaeda [39].

2.7.3 Dynamic Network Analysis

People do understand the hierarchies of typical organization but they may not be familiar with the hierarchies of the convert or terrorist organization as they possess a different structure. One of the key characteristic of such networks is to be distributed and cellular. So, the dynamic nature of such networks demands to treat them as a special case not like the ordinary network. The Dynamic Network Analysis (DNA) concept was presented based on the above discussed features in [40]. To cope with the dynamic nature of such covert networks, DNA was introduced by extending the traditional measures. The key advancements of DNA are as follow:

- The Meta-Matrix
- Relations are treated as variables which gives the ability to have related weight or probability
- Combination of cognitive science and social networks or of social networks with multi agent system, which prompt the agents to adapt

The ideology behind this method to be proposed is to analyze the performance of the system to calculate the degree of destabilization. This is done by calculating system performance, finding key player using any centrality measure, remove them and again check the performance. Data of the bombing on an embassy in Tanzania was used to test DNA as a proof of concept.

2.7.4 Investigative Data Mining

The ability to map a hidden cell, to calculate the particular structural and interactional criteria of a hidden cell are offered by Investigative Data Mining (IDM) [28]. The IDM focuses on to discovering the patterns of individual's relations, analyzing discovered patterns and on basis of this predicting the behavior and decision making of the network. IDM also helps us to calculate the efficiency of the hidden cell and level of secrecy. It also helps us to measure the activity level, access ability for nodes and the control level on the hidden cell. These features enable to develop applications for counter terrorism, which will be helpful in understanding and defusing a terrorist network. Studies tells the multiple investigative data mining approaches. One of such approach is link analysis which focuses on finding the structural features of network using search and probabilistic techniques [41]. The structural features are like gatekeepers, hubs, pulse takers and also identification of relationships. To find the hidden relationships in the terrorist network link analysis approach was proposed and implemented. In [28] four case studies and involved network have been analyzed. The selected case studies are Bali Night Club Bombing Terrorist Attack, WTC 1993 Bombing Plot, Dirty Bomb Plot and September 11, 2001 Terrorist Plot.

Chapter 3

Proposed Model

Chapter 3

3.1 Introduction

In this chapter we discuss the proposed methodology for finding correlation or similarity of conversations between different nodes and data dictionary and also among different nodes as well to detect the key players using text mining in detail. Current methods of social network analysis mainly focus on the different degree measures considering nodes and edges of the network, while there is still a lot of work need to be done keeping the focus on the communication between nodes. As mentioned in relative degree in related work, a node in terrorist network might not be prominent but that particular node might be the leader of the network. Group leaders of terrorist network possess some distinct characteristics concluded from the observation of actual terrorist networks datasets. So, there might be a chance that group leaders communicate very rarely but to the point, using code words and to some most influential members of the network. Removing that node from the network will help in destabilizing the network. But, what are they planning as an organization, we are still not able to extract that from the existing methods of social network analysis. In this research we mainly focus on finding the correlation of conversation with data dictionary and among different nodes in network.

3.2 Text mining process

Figure 3.1 shows a high level process for text mining, which remain same across all sort of text mining techniques. This figure outlines the three main component of text mining. First is to gather the data from different sources, data is always in unstructured form. Data can be gathered from any source like documents, emails, web pages etc. Second component used to transform gathered unstructured data into structured meaningful information for further use, it is usually named as preprocessing. While third component is responsible to apply any sort of text mining algorithms or techniques to predict something, analysis or what else user required. These three component grouped together to form a useful text mining application, which could help in handling large unstructured data.

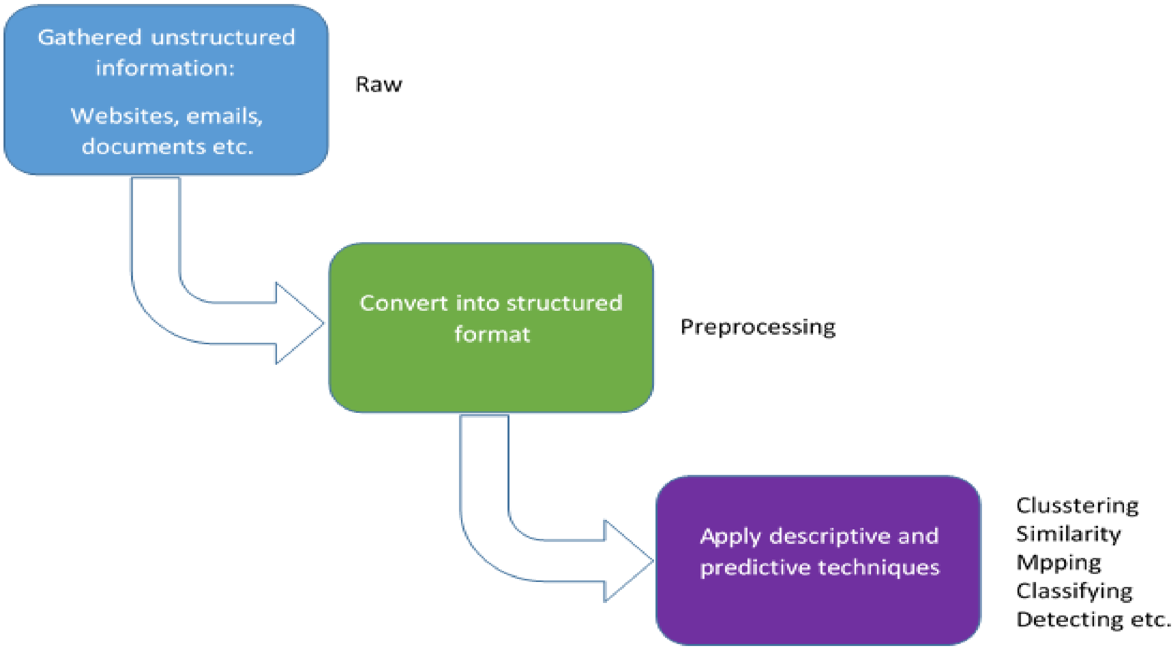

Gathered unstructured information:
Websites, emails, documents etc.
Raw
Convert into structured format
Preprocessing
Apply descriptive and predictive techniques
Clusstering
Similarity
Mpping
Classifying
Detecting etc.

Figure 3.1 High level general text mining process

3.3 Architecture

The main purpose of our proposed framework is to detect the key player from the terrorist network by applying the text mining techniques on communication of nodes. It is a new aspect in social network analysis. Figure 3.2 shows the architecture of the framework. The module "Loading communication data and dictionary" gets the structured communication of nodes as well as data dictionary for further action on them in the framework. Module "Process Selection" gives us a choice to generate a new process or to load an already saved process, if key player detection is already done and needs to be extend or modify for more nodes. "Generate words vector" module is for generating word vectors for every node's communication and data dictionary, so similarity between them could be found. This module mostly used the preprocessing steps. Next comes the "Correlation Calculation' module which gets the results in words vector form and calculate the similarity between data dictionary and selected nodes. In last evaluation and analysis of results carry out to detect the key player. We can save the process as well as its results.

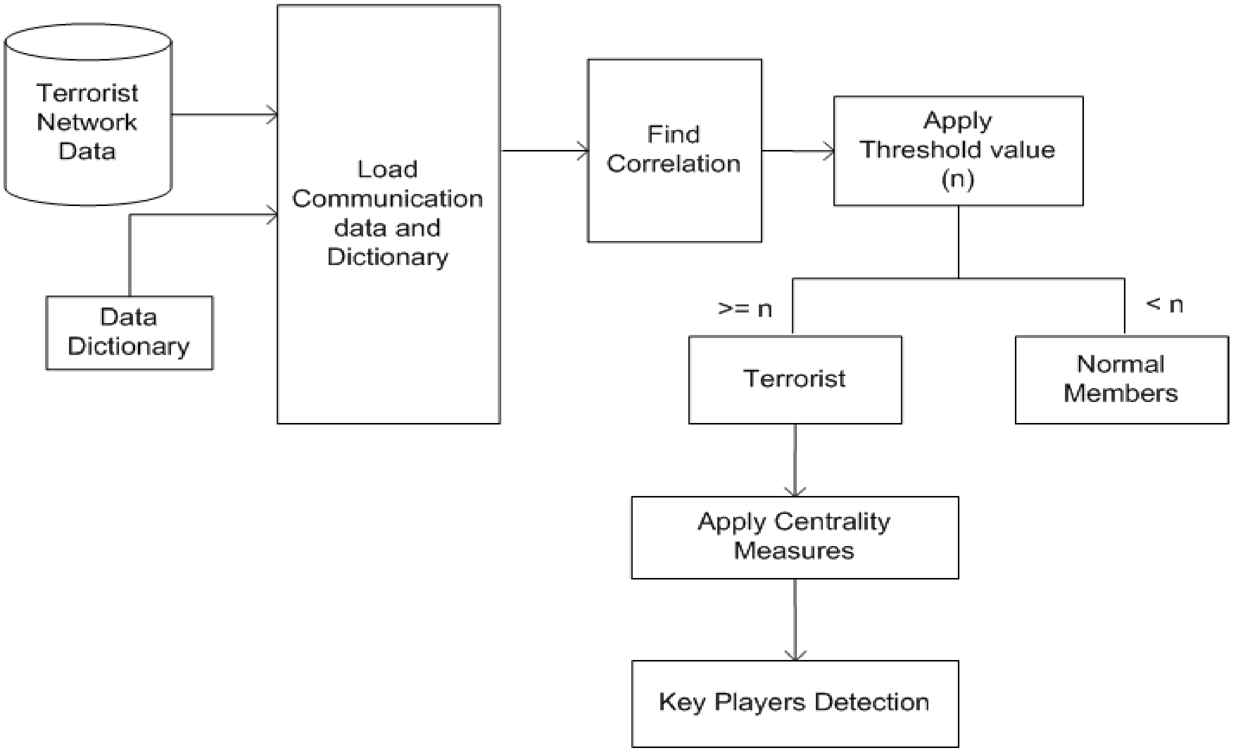


Figure 3.2 Architecture diagram of key player detection framework using communication

3.4 Components of proposed model

Proposed model for key player detection is basically consists of three parts. First part or a pre requisite to have a data dictionary which must contain all the words and code words, which terrorist use in their communication. (This part might be optional if there is a pre-defined dictionary available of terrorist communication). We developed a words dictionary from the available dataset. Second part of the model is preprocessing of the given data, for a particular node or nodes at a time and storing them as csv document or in other format for further use in the process. Third part of the model is finding the similarity/correlation between pre-processed data and pre-defined dictionary. After that analysis of results and evaluation of key player detection carries out. On basis of the analysis and evaluation we can conclude whether a node is key player or not in the network. Text mining to detect the key players, is a novel approach of social network analysis for convert networks.

3.4.1 Data Dictionary

Data dictionaries consists of descriptions of data items or objects of a data model for developers or user to use them as reference. It is mandatory to have a data dictionary to analyze any system. We used words dictionary as a data dictionary in this thesis. Words dictionary consists of words used by terrorist in their conversation, which may include code words which they use specifically among their organization to make plans for different events or use in their communication.

For reference we develop a words dictionary using the available communication data set. We fetched different conversation chunks of different nodes to develop this words dictionary. To develop the words dictionary different steps was taken.

- Merging of all conversation chunks
- By using text mining, tokenizing the words and saving words vector to a document as words

3.4.2 Data Preprocessing

Data preprocessing is the one important and integral part of the proposed model. A labelled data set was required for our proposed model. This data set is used to test and train our model. As a labelled data set, we choose Enron email data as a test dataset which is publically available for research works. Terrorist network data set with labelled communication was not available at ease due to its sensitivity. Pre-processing of data means to labeled the data, remove anomalies, convert it to a structured form, remove duplicate values, convert words to same case and convert words to their base value and then transform it into word vector to find the correlation or similarity. This also includes to write the pre-processed data into a csv file at desired location. Figure 3.3 shows the preprocessing design and work flow. Using preprocessing we also generate the data dictionary, because we do not have a pre-defined dictionary, so we decided to create our own from the communication of the nodes.

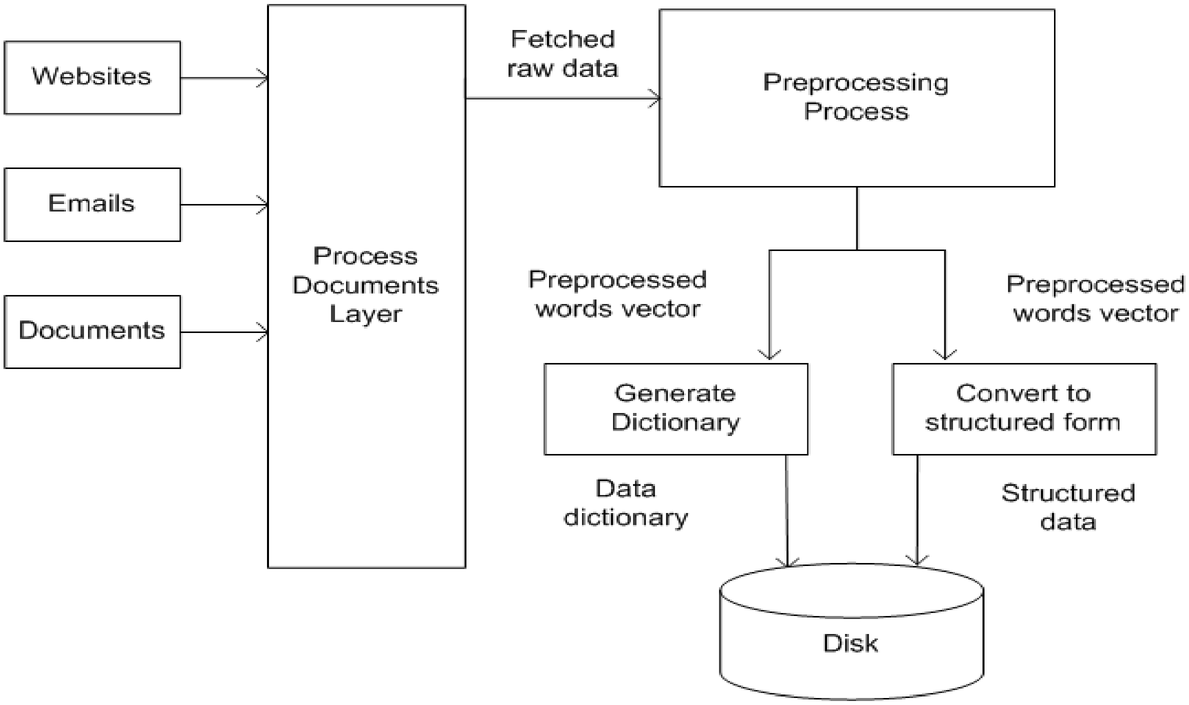


Figure 3.3 Preprocessing of the data

Operators used for Preprocessing

Different operators are used for the preprocessing in rapid miner, which provides numerous text mining extensions. A brief detail of each operator used is given below:

Process documents from files: Text mining is used in order to do the preprocessing. For this purpose we used "Process Documents from Files" operator in rapid miner. This operator is used to output a word vector from text fetched from a single or multiple files. Communication data of any node could be in one file or in multiple files. As in our data set it consists of multiples files and from multiple folders as well. So, we used this operator to fetched data from different files for a node and after doing applying preprocessing on them to convert in structured form, we write the output in one file labeled accordingly. This help out us in structuring the unstructured data as well.

Then we move into sub process window by double clicking the "Process document from files" operator, to perform the actual text mining. This is the place where multiple operator are linked to take the selected document and breaks down them into words.

In sub process window we used following operators to extracts words from different documents:

- Tokenization
- Filter Tokens (by length)
- Filter Stopwords (English)
- Stemming (Porter)
- Transform Cases

Tokenization: “Tokenization is the process of breaking a stream of text up into phrases, words, symbols, or other meaningful elements called tokens”. To extract the words in a sentence in the main purpose of the tokenization. In text mining, for information retrieval from given text, words of that text or data set are required, which demands of the parser for tokenization of documents. The purpose of tokenization is also the identification of significant words.

Filter Tokens (by length): Filtering provides a great deal of elasticity while designing of comprehensive data sources and mining structure. It helps to build a one mining structure if you have a well understood view of data source, for different models. We set the minimum and maximum characters for token lengths to 2 and 25 respectively.

Filter Stopwords: Prepositions, pronouns and articles are such words which are not going to be helpful in text mining. And these words are known as stop words in text mining, so these are helpless in text mining process. By eliminating all such words will help in results precision of text mining. This will also minimize text data, which in turn will improve the performance of the system. We used “**Filter Stopwords** (English)” operator for this purpose.

Stem (Porter): To transform words into their base word, root or stems is called stemming. For example transforming liking or likely to like is called stemming. Stemming is quite useful in text mining because all words are transformed in to their same root. We used “**Stem** (Porter)” operator for this purpose.

Transform Cases: Transform cases is the process of transforming all extracted words in to lower case or upper case for symmetry. We used “**Transform Cases**” operator for this purpose. We used the option to convert all words in lower cases. Same case will help in symmetry of the words, eliminating the confusion factor for same word, if exist in both cases. Figure 3.4 shows the connected operator for data preprocessing.

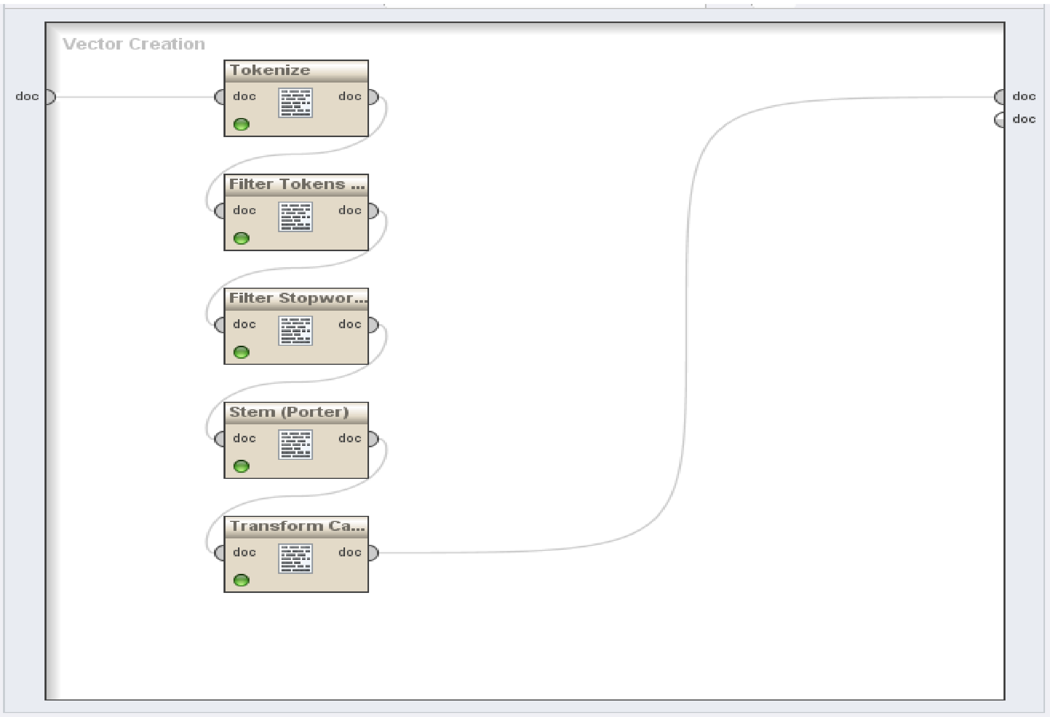


Figure 3.4 Sub-process window to show pre-processing operator

WordList to Data: After performing the preprocessing on the fetched documents, we move back to main process window, where results are now passed to another operator named “WordList to Data” from the process document from file. This operator transform a word list into a dataset. The data set consists of words and their attributes, which includes number of documents in which the words occurred, labelled document, its class. This operator is basically used to filter the wordlist before to be used for any other purpose.

Write CSV: After preprocessing and transforming a wordlist to dataset, we need to save the result to be further used in finding correlation or similarity. For this purpose and to avoid the re-preprocessing of the same data, we write the output the results into csv files named after the node’s label. We used “Write CSV” of rapid miner for this purpose.

Figure 3.5 shows the converting the unstructured and un-organized data to structured form and saving in csv file.

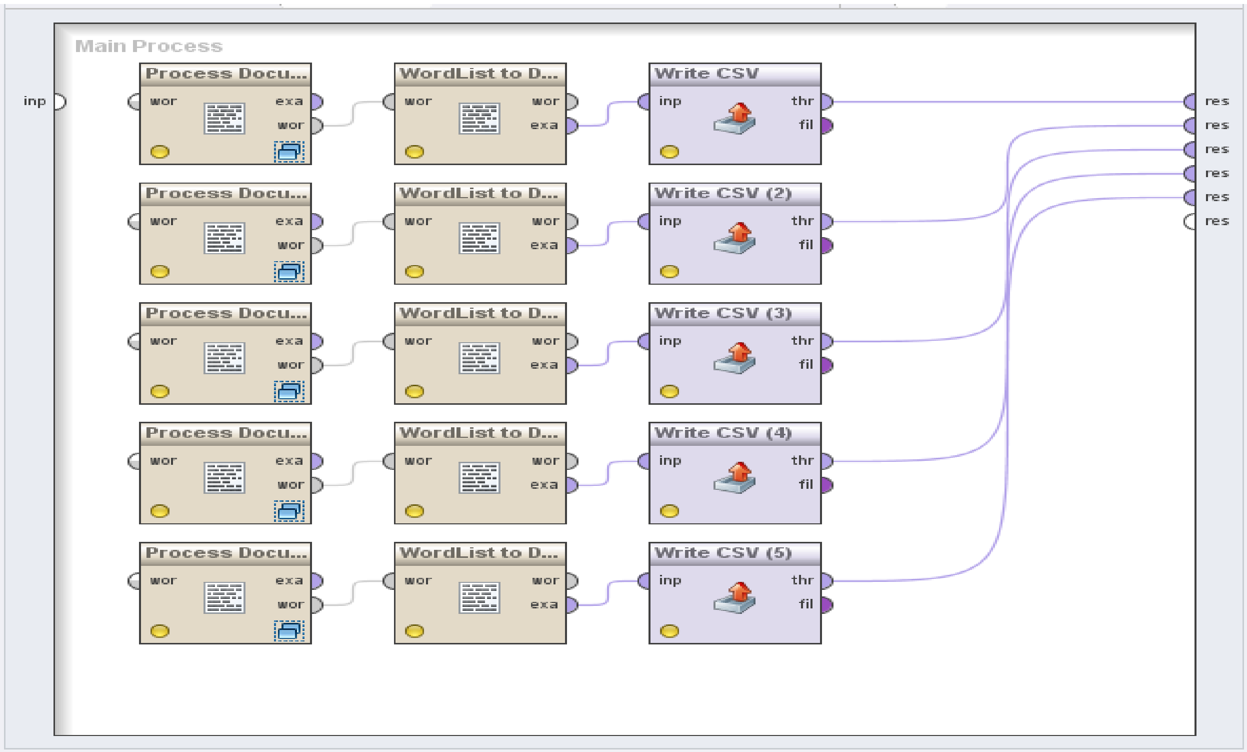


Figure 3.5 Process to convert unstructured data to structured data

This preprocessing could be applied on the data of one node or on the data of multiple nodes at the same time. But the best practice would be to follow the relationship a suspect node has in a network. For example if there are six links going out/in to a node, communication of all six links can be preprocessed at the same time and a separate CSV file would be generated for each link or node. As, the results of preprocessing are written in a file, this could be beneficial in future use if same node appears in connection with any other suspected node, no further preprocessing required at that time, we can consume that file again.

3.4.3 Similarity Model

Overview: This is the third part of proposed model. This process also consists of three parts itself. As in preprocessing we structured the unstructured data of all nodes in the form of a separate csv file for each node. Every csv file now contains words of all communication of one node. For finding the similarity, we again used the preprocessing step, but this time not for the unstructured data but for csv file of a node with other csv files, to extract the word vectors of different nodes. Here, we only use partial preprocessing, we do not write csv files in this step.

After “Process documents from files”, we use “Data to similarity” operator, which calculate the similarity between data dictionary and different selected nodes. Similarity between documents returns a matrix which shows the degree of similarity between data dictionary and node, or between data dictionary and each node respectively if multiple node’s communication is selected for correlation calculation. By the results we could conclude the key players if their similarity is at higher side.

To find the correlation or similarity, first we need to detect the suspect node and then the linked nodes in terrorist network. This could help us to run our model on the only selected s nodes, so that it becomes easy to find the bonding between linked nodes in network. Benefit of labelled data set was that we could generate the structured data using node’s identification and then consume that structured data in following preprocessing step. After detecting some nodes in network, we fetches their structured conversation in the following preprocessing step along with the data dictionary to find the correlation.

Preprocessing: At this level preprocessing is performed to get the words vector to find the similarity of each node with the predefined data dictionary. Here we have the provision to find the correlation of conversations among nodes or similarity of conversations with predefined data dictionary. But in our proposed model, we find correlation of each with data dictionary. We detect a group of suspected or linked nodes, and fetches their structured communication, also data dictionary is selected and apply preprocessing on them.

Data to Similarity: After preprocessing of the data of selected nodes with data dictionary, we need to find the similarity or correlation between conversations of different nodes and data dictionary. Finding correlation or similarity between pre-defined dictionary and network node means that how much communication of that specific node match with the pre-defined dictionary. A high score of similarity shows that a node is using mostly those words which lie in the pre-defined dictionary and low score of similarity shows that a node does not used words from data dictionary or use very less words in its conversation. This predicts that a node with less similarity score is unaware of the terrorist in the network and it has no link with them. While high score of similarity shows that a node is fully aware about data dictionary, words, code words and when to use them. It definitely have links with terrorist or it may be the terrorist. As normal node do not know about those code words used by the terrorist, so whoever is using those words in its communication, shows s/he belongs to any terrorist group? We can predict based on our correlation/similarity model that whether a node is terrorist or not, and if yes then it is a low level terrorist or lie in the category of key players of terrorist.

What is similarity? A text document could be modeled in different ways, for example it could be broken down into words and it could be shown as carrier of words. Here, order of words do not matters and words don't have any dependency. This bag of words model is quite popular in text mining and retrieving information. One important aspect of representation is that each word in carrier resembles to an aspect in that data space and that leads to the vector representation of that document containing the positive values on each dimension.

For example, we have a document set $D = \{d1, \dots \dots, dn\}$n} and a set of different terms which exist in D like $T = \{t1, \dots \dots, tm\}$. Here the term is referring to a word. So a document or bag of words could have represented as m-dimensional vector td. If a term $\boldsymbol{t} \in \boldsymbol{T}$ of a document $\boldsymbol{d} \in \boldsymbol{D}$ then $tf(d, t)$ will refer to the term frequency. This will lead us to the document's vector representation like:

$$\overrightarrow{t_d} = (tf(d, t1), \dots \dots, tf(d, tm)) \qquad \textbf{(3.1)}$$

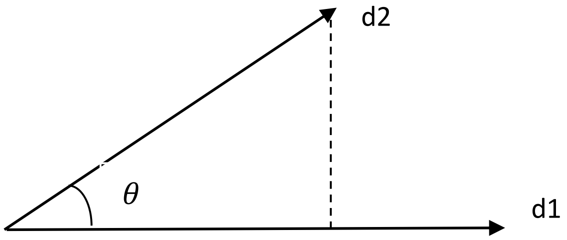


Figure 3.6 Cosine similarity angle between two vectors

Above diagram depicts the documents representation in vector form. So, similarity between two documents d1 and d2 is being calculated as the correlation between the vectors of those documents, the result of that correlation would be the cosine of the angle between respective vectors.

There are always a strong case that most frequents words in a vector might be not the most informative and important. On the other hand words that are occurring regularly in some documents but seldom in other documents, lean towards to the more appropriate for the outcome of similarity. In our case that might be the word from the data dictionary. Such words are more appropriate to find the similarity. To fetch such words, the above defined term frequency has to be changed and we do it by transforming that into a weighting scheme named tfidf (term frequency and inversed document frequency). In a document d the term frequency is weighted by tfidf, it counts the factor that does not count its occurrences in the collection of documents for its importance. Tfidf is defined as:

$$tfidf(d,t) = tf(d,t) \times \log(\frac{|\mathrm{D}|}{df(t)}). \qquad \textbf{(3.2)}$$

In above mentioned equation, df(t) represent the no. of documents in which word occurs. We used tfidf to generate the term vector rather than the total term frequency.

Similarity Measures: There are many similarity measures in our selected operator “Data to similarity”. These similarity measure depicts the closeness of one word vector with others. The selection of the similarity measure mostly depends on the nature of data and its characteristics, reason being that no one similarity measure is best suitable for all data types.

Furthermore, selection of suitable similarity measure is also very important and vital for a particular nature of data. To make such selections, one must have to understand the similarity measures in detail and their usefulness too. There are four types of measurements are available for “Data to similarity” operator in rapid miner. Then each type contains different measures itself.

Following are the types of measurements are available:

- Mixed measures
- Nominal measures
- Numerical measures
- Divergence measures

Each of the above similarity measure has different similarity measures, each similarity measure works for the data which have some particular characteristics. Each similarity measure has some differences in working from other measures.

For our proposed model, we need to select a similarity measure from the above mentioned measures, which is best suitable to find correlation between vectors of words, and in result give some score, on basis of that we could conclude something to detect key player from network. So, for our model we have designed some pre-requisites to choose the similarity measure. Following are the criteria used for selecting the similarity measure:

- Similarity measure should be capable of working on words
- Its result should be easily understandable and in numeric form (and good if in matrix form)
- It could be efficient
- It performs similarity of one vector only once with others, not vice versa.

Based on above criteria, we choose the "cosine similarity" measure of "Numerical Measures" for our proposed model. This is one of the most popular similarity measure applied on text documents.

Cosine Similarity: When the documents are tokenized in the form of term or word vectors, then similarity between two documents corresponds to correlation between vectors for such representation. And it is reckoned as the "cosine of the angle" between the respective vectors, this measurement is also known as the cosine similarity. In context of the text documents mining, this similarity measure is known as one of the most famous to be applied on the text documents, for example it is mostly used in many information retrieval applications, further more in clustering too. Cosine similarity between two documents $\overrightarrow{t_a}\ and\ \overrightarrow{t_b}$ can be computed as:

$$SIMc\left(\overrightarrow{t_a}, \overrightarrow{t_b}\right) = \frac{\overrightarrow{t_a} . \overrightarrow{t_b}}{|\overrightarrow{t_a}| \times |\overrightarrow{t_b}|} \qquad \textbf{(3.3)}$$

In above equation $\overrightarrow{t_a}\ and\ \overrightarrow{t_b}$ are two m-dimensional vectors representation of respective documents above the words set $T = \{t_1, \dots\dots, t_m\}$. A dimension is the representation of the term having a non-negative weight in the respective document. This leads the cosine similarity between two documents to non-negative values and it is confined between [0, 1].

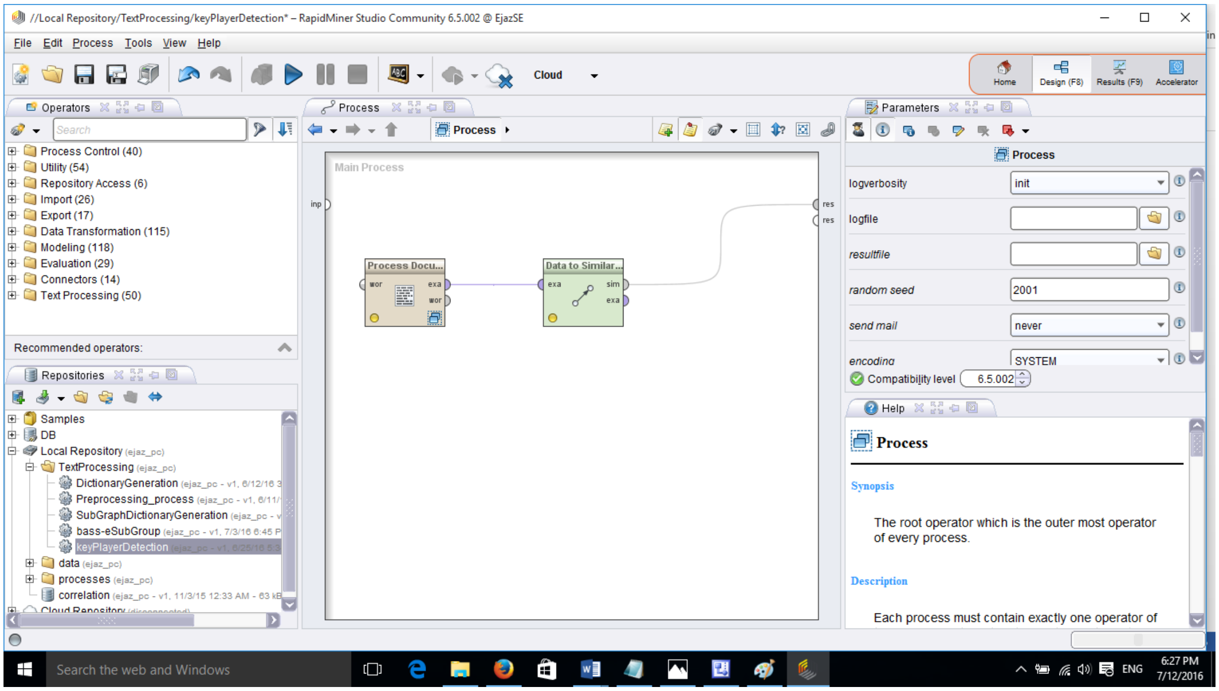
One of the most important characteristic of this similarity measure is that, it is not dependent on the document length. For an instance, a document d' is obtained after joining the two same copies of the document d, then computing the similarity between these two documents will return 1, depicting that the documents used for similarity computation are observed to be same. In the meanwhile, if there are three documents are given labelled *d*, d' and *l* then their similarity score will also be equal to 1, like $sim\left(\overrightarrow{t_d}, \overrightarrow{t_l}\right) = sim\left(\overrightarrow{t_{d'}}, \overrightarrow{t_l}\right)$. So this shows, documents will be considered same if they contain the same contents but in total are diverse. Although, by joining the two replicas of document a different document is obtained, this object is not similar to the original document in general, but cosine similarity score shows, this leads to above description not to meet the second rule of the metric.

Model Implementation: Once pre-processing of desired nodes are done and we have a data dictionary, then comes finding the correlation between communication of nodes and pre-defined dictionary. We use “Data to similarity” operator of rapid miner for this purpose. On basis of correlation or similarity score which would be bounded to [0, 1], we’ll decide about the key players in the network. There two main operator to carry out this process. One is “Process document form files” and second is “Data to similarity” operator. Purposes and working of both operators are described in detail in above sections.

To perform this process, we first put the “Process document form files” in main process window from operators tab. In sub-process window, we apply the above mentioned pre-processing operators to construct the term vectors of the respective document of suspected nodes selected in the parent operator in main process window. It is need to be remember that data dictionary also selected as a document as well. This operator will output all the term vectors. After that we select the “Data to similarity” operator from the operators tab and put it in main process window. Then we connect the exampleSet port to the input port of the “Data to similarity” operator, which takes exampleSet as the input. Then, we select the similarity measure type given in the parameter tab. Here four options are available, these are mixed measures, nominal measures, numerical measures and divergence measures. We select the numerical measures. There are also multiple similarity measure available in this type. We choose the cosine similarity from given similarity measures. Then we connect the **sim** output port to the result port. By pressing the run button, process run and after finishing it displays the results.

Here, we have the option of how we want to view the result of correlation, in matrix form or in one to one correspondence. For one to one correspondence, we connect the sim and exp ports with result ports of main process and run the process. Then it shows the results. One to one correspondence is preferred if similarity needs to be calculated between data dictionary and network nodes.

Figure 3.7 shows the correlation process.



//Local Repository/TextProcessing/keyPlayerDetection* – RapidMiner Studio Community 6.5.002 @ EjazSE

Process

logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

Compatibility level 6.5.002

Help

Process

Synopsis

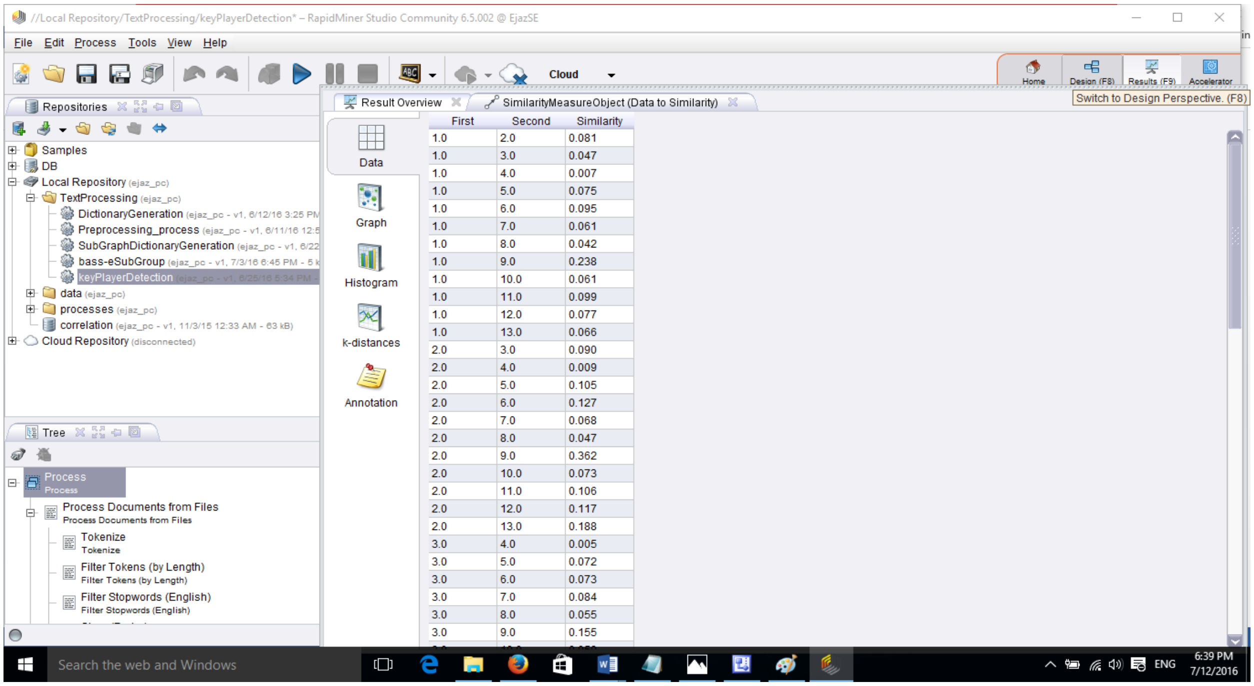
The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of

Figure 3.7 Correlation process

Figure 3.8 shows the results of the above mentioned process.



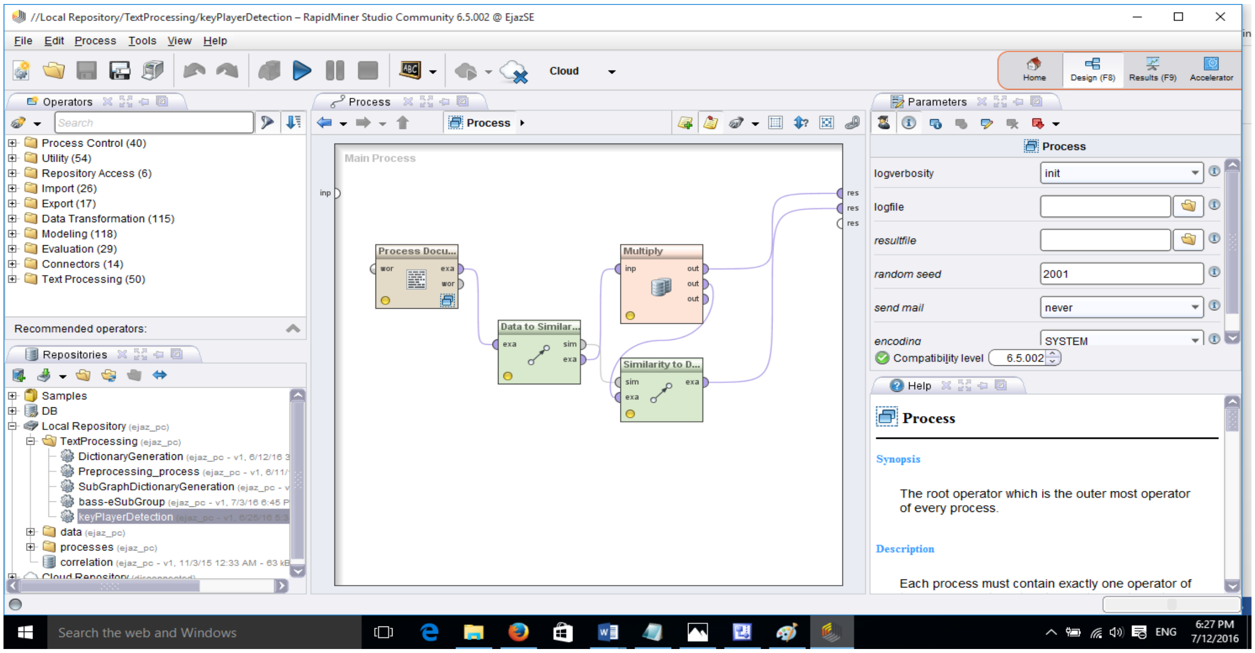
//Local Repository/TextProcessing/keyPlayerDetection* – RapidMiner Studio Community 6.5.002 @ EjazSE

File Edit Process Tools View Help
Cloud
Home
Design (F8)
Results (F9)
Accelerator
Switch to Design Perspective. (F8)
Repositories
Samples
DB
Local Repository (ejaz_pc)
TextProcessing (ejaz_pc)
DictionaryGeneration
Preprocessing_process
SubGraphDictionaryGeneration
bass-eSubGroup
keyPlayerDetection
data (ejaz_pc)
processes (ejaz_pc)
correlation (ejaz_pc - v1, 11/3/15 12:33 AM - 63 kB)
Cloud Repository (disconnected)
Tree
Process
Process Documents from Files
Tokenize
Filter Tokens (by Length)
Filter Stopwords (English)
Result Overview
SimilarityMeasureObject (Data to Similarity)
Data
Graph
Histogram
k-distances
Annotation
Search the web and Windows
ENG
6:39 PM
7/12/2016

First	Second	Similarity
1.0	2.0	0.081
1.0	3.0	0.047
1.0	4.0	0.007
1.0	5.0	0.075
1.0	6.0	0.095
1.0	7.0	0.061
1.0	8.0	0.042
1.0	9.0	0.238
1.0	10.0	0.061
1.0	11.0	0.099
1.0	12.0	0.077
1.0	13.0	0.066
2.0	3.0	0.090
2.0	4.0	0.009
2.0	5.0	0.105
2.0	6.0	0.127
2.0	7.0	0.068
2.0	8.0	0.047
2.0	9.0	0.362
2.0	10.0	0.073
2.0	11.0	0.106
2.0	12.0	0.117
2.0	13.0	0.188
3.0	4.0	0.005
3.0	5.0	0.072
3.0	6.0	0.073
3.0	7.0	0.084
3.0	8.0	0.055
3.0	9.0	0.155

Figure 3.8 Results in One to One Correspondence

But if similarity needs to be calculated between selected nodes, then it would be better to view the results in matrix form. For this purpose we need to split the output with “Multiply” operator. Connect the ‘exa’ port of data to similarity operator with input of “Multiply”. Here, we also put another operator named “Similarity to Data” in main process window. The ‘sim’ output port of “Data to similarity” operator is now connected to the ‘sim’ input port of the “Similarity to Data”. We need to set a parameter for “Similarity to Data” operator named “table type”. We select the value matrix. Then we connect the one output port of “Multiply” operator with ‘res’ of main process and another output port with the ‘exa’ input port of the “Similarity to Data”. In last we connect the ‘exa’ output port of the “Similarity to Data” operator with main process window output port and run the process. Figure 3.8 shows the process that displays the results in matrix form.



logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

Compatibility level 6.5.002

Process

Synopsis

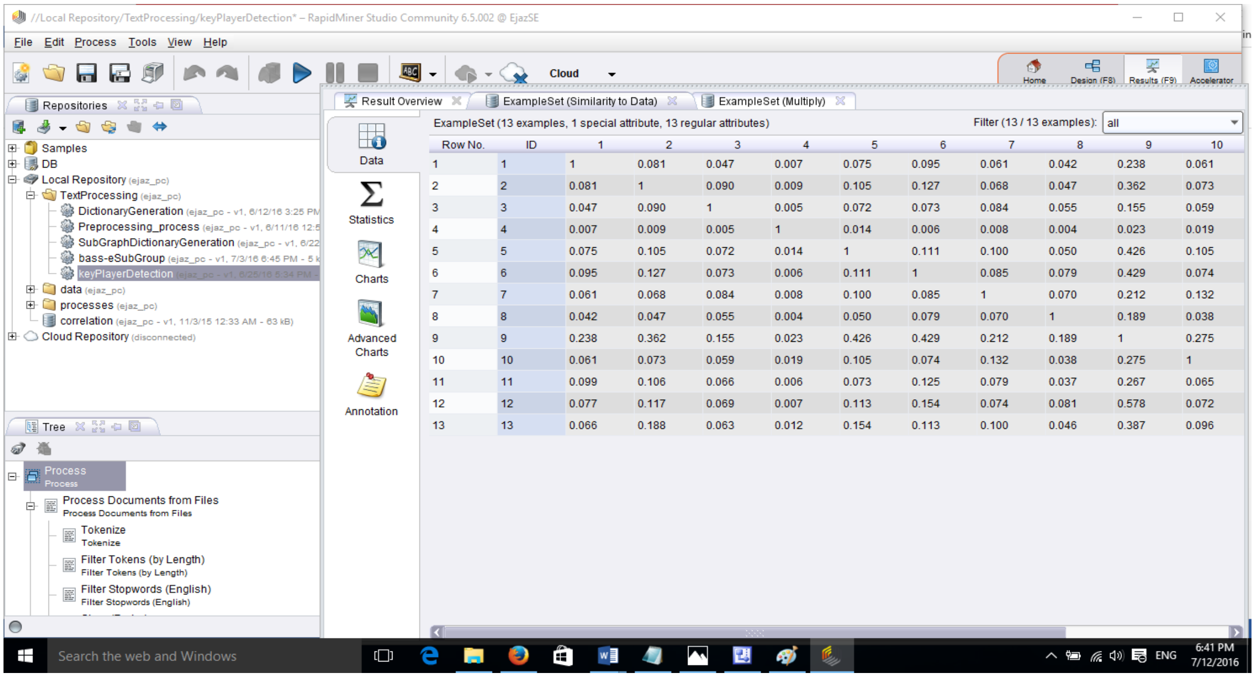
The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of

Figure 3.9 Correlation process that displays results in matrix form

Figure 3.10 show the results of the correlation in matrix form.



//Local Repository/TextProcessing/keyPlayerDetection* – RapidMiner Studio Community 6.5.002 @ EjazSE

ExampleSet (13 examples, 1 special attribute, 13 regular attributes)

Filter (13 / 13 examples): all

Row No.	ID	1	2	3	4	5	6	7	8	9	10
1	1	1	0.081	0.047	0.007	0.075	0.095	0.061	0.042	0.238	0.061
2	2	0.081	1	0.090	0.009	0.105	0.127	0.068	0.047	0.362	0.073
3	3	0.047	0.090	1	0.005	0.072	0.073	0.084	0.055	0.155	0.059
4	4	0.007	0.009	0.005	1	0.014	0.006	0.008	0.004	0.023	0.019
5	5	0.075	0.105	0.072	0.014	1	0.111	0.100	0.050	0.426	0.105
6	6	0.095	0.127	0.073	0.006	0.111	1	0.085	0.079	0.429	0.074
7	7	0.061	0.068	0.084	0.008	0.100	0.085	1	0.070	0.212	0.132
8	8	0.042	0.047	0.055	0.004	0.050	0.079	0.070	1	0.189	0.038
9	9	0.238	0.362	0.155	0.023	0.426	0.429	0.212	0.189	1	0.275
10	10	0.061	0.073	0.059	0.019	0.105	0.074	0.132	0.038	0.275	1
11	11	0.099	0.106	0.066	0.006	0.073	0.125	0.079	0.037	0.267	0.065
12	12	0.077	0.117	0.069	0.007	0.113	0.154	0.074	0.081	0.578	0.072
13	13	0.066	0.188	0.063	0.012	0.154	0.113	0.100	0.046	0.387	0.096

File
Edit
Process
Tools
View
Help
Cloud
Home
Design (F8)
Results (F9)
Accelerator
Repositories
Samples
DB
Local Repository (ejaz_pc)
TextProcessing (ejaz_pc)
DictionaryGeneration (ejaz_pc - v1, 6/12/16 3:25 PM
Preprocessing_process (ejaz_pc - v1, 6/11/16 12:5
SubGraphDictionaryGeneration (ejaz_pc - v1, 6/22
bass-eSubGroup (ejaz_pc - v1, 7/3/16 6:45 PM - 5 k
keyPlayerDetection (ejaz_pc - v1, 6/25/16 5:34 PM -
data (ejaz_pc)
processes (ejaz_pc)
correlation (ejaz_pc - v1, 11/3/15 12:33 AM - 63 kB)
Cloud Repository (disconnected)
Tree
Process
Process
Process Documents from Files
Process Documents from Files
Tokenize
Tokenize
Filter Tokens (by Length)
Filter Tokens (by Length)
Filter Stopwords (English)
Filter Stopwords (English)
Result Overview
ExampleSet (Similarity to Data)
ExampleSet (Multiply)
Data
Statistics
Charts
Advanced Charts
Annotation
Search the web and Windows
ENG
6:41 PM
7/12/2016

Figure 3.10 Correlation process that displays results in matrix form

Chapter 4

Result and Key Player Detection

Chapter 4

4.1 Overview

After discussing the feature of model and its architecture to find the correlation between the data dictionary and communication of network's node in previous chapter, we discussed and analyzed the result obtained from the model and how the model and results could help us to detect the key player in terrorist network. We discussed the key player detection from results and also the how the different nodes have relationship among themselves in the said network.

4.2 Modeling Tools

Different tools like Rapidminer studio 6.5, NodeXL, Microsoft Excel, MS visio etc are used in the research work for different tasks. Rapid miner studio is used for building the model, applying the model, obtaining the results and saving the model to be used in future with little modification in the developed process. It also used for some charts and graphs representation of the results. NodeXL is mainly used for the graphical representation of the chosen Enron data set and also for understanding the different measures of a social network to kick start off our research work. Form that network we select a node and its relationship nodes to draw a case study which will be used to prove the concept of our model and finding the results. Microsoft excel is used for some analysis and graphical representation of results. MS visio is used for the different system design diagrams.

4.3 Enron network

We choose the Enron public email dataset as an example dataset for our research work. The main reason for choosing this dataset was, because it provides the communication data between network nodes and also data to construct the actual network as well. Figure 4.1 shows the enron network of 150 users, which communicate through emails. These users are mostly senior officials. Figure also shows some highlighted edges in red color and also a node named allen-p highlighted. This depicts that node allen-p is selected and all its linked edges also get selected. We construct this network using NodeXL providing the Enron network information telling nodes and relationship among them.

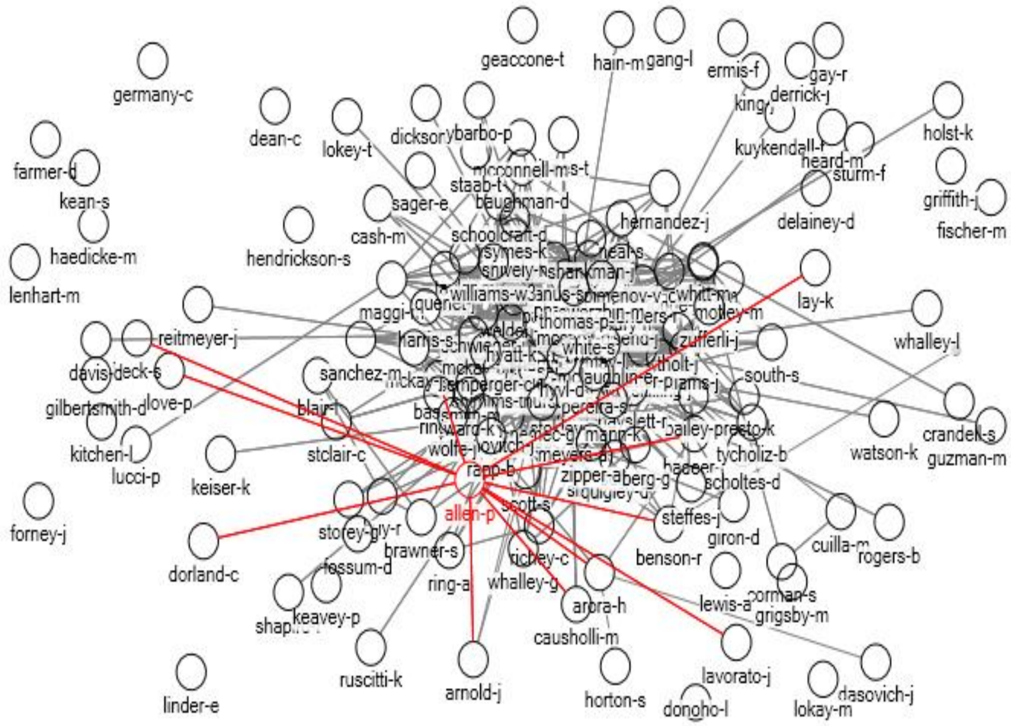


Figure 4.1 ENRON network diagram

4.4 Features of proposed model

Our model could work in different ways. It could operate to find the similarity between data dictionary and a suspicious node alone, it could work on different nodes to find correlation among themselves or between nodes and data dictionary simultaneously. This flexibility provides us the option of how we want to manipulate networks nodes. This also provides that we could find the same nature of nodes in a network as well. So there are multiple ways of operating with the model for social network analysis to find the key player based on their communication. These options includes:

- Finding correlation between a single node and data dictionary
- Finding correlation between multiple nodes and data dictionary at same time
- Finding correlation among multiple nodes at same time

Above three options provide the flexibility to operate the model, depending on how user want to analyze the network. Also, there is ease to use above mentioned options in the model, because user does not need any extra steps or measure to take for finding the similarity or correlation.

In this research we provide the results of two methods. One is for detecting key player by finding the correlation between data dictionary and different nodes which is primarily requirement of the research and also correlation among the multiple nodes as well.

4.5 Case study

For finding the correlation we choose a suspicious node in our case say it is labelled "allen-p" in network. Then we find all the linked nodes with selected node. Figure 4.2 shows the sub-network of all nodes linked to "allen-p" node.

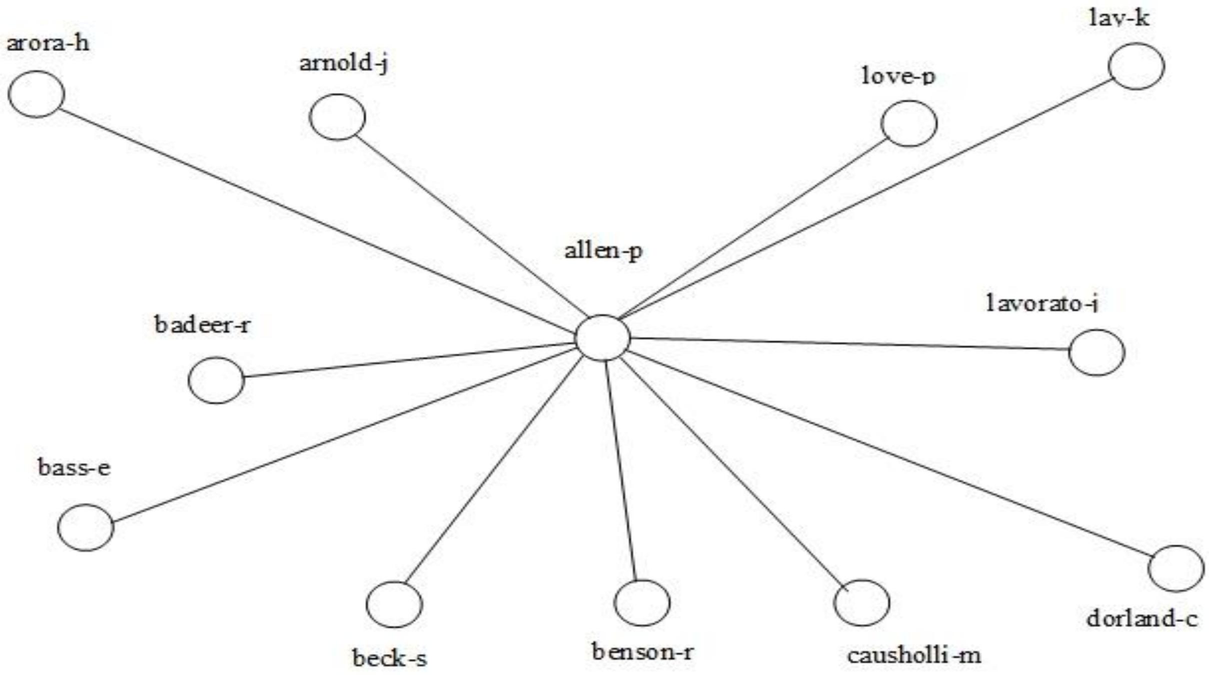
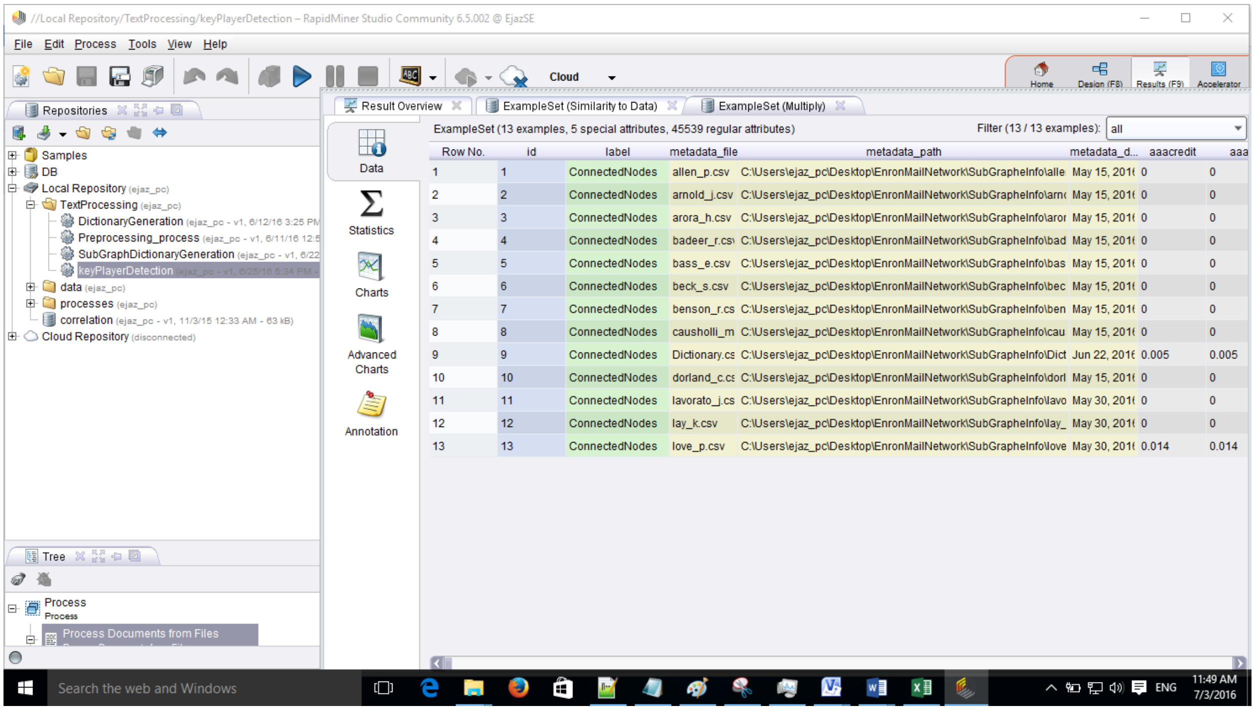


Figure 4.2 – Suspicious node with linked nodes

Now, we need to find the correlation of all nodes linked to "allen-p" including "allen-p" with data dictionary, which will shows which node is using the most words from data dictionary enabling us to conclude and detect the terrorist or key nodes in network. Figure 4.3 shows the output of example set after finding the similarity. This shows the label of all selected nodes and their respective ids which rapid miner internally assigns them all. By using these ids we can analyze the results of similarity scores.


//Local Repository/TextProcessing/keyPlayerDetection – RapidMiner Studio Community 6.5.002 @ EjazSE
File
Edit
Process
Tools
View
Help
Cloud
Home
Design (F8)
Results (F9)
Accelerator
Repositories
Samples
DB
Local Repository (ejaz_pc)
TextProcessing (ejaz_pc)
DictionaryGeneration (ejaz_pc - v1, 6/12/16 3:25 PM
Preprocessing_process (ejaz_pc - v1, 6/11/16 12:5
SubGraphDictionaryGeneration (ejaz_pc - v1, 6/22
keyPlayerDetection (ejaz_pc - v1, 6/25/16 5:34 PM -
data (ejaz_pc)
processes (ejaz_pc)
correlation (ejaz_pc - v1, 11/3/15 12:33 AM - 63 kB)
Cloud Repository (disconnected)
Tree
Process
Process
Process Documents from Files
Result Overview
ExampleSet (Similarity to Data)
ExampleSet (Multiply)
Data
Statistics
Charts
Advanced Charts
Annotation
Search the web and Windows
ENG
11:49 AM
7/3/2016

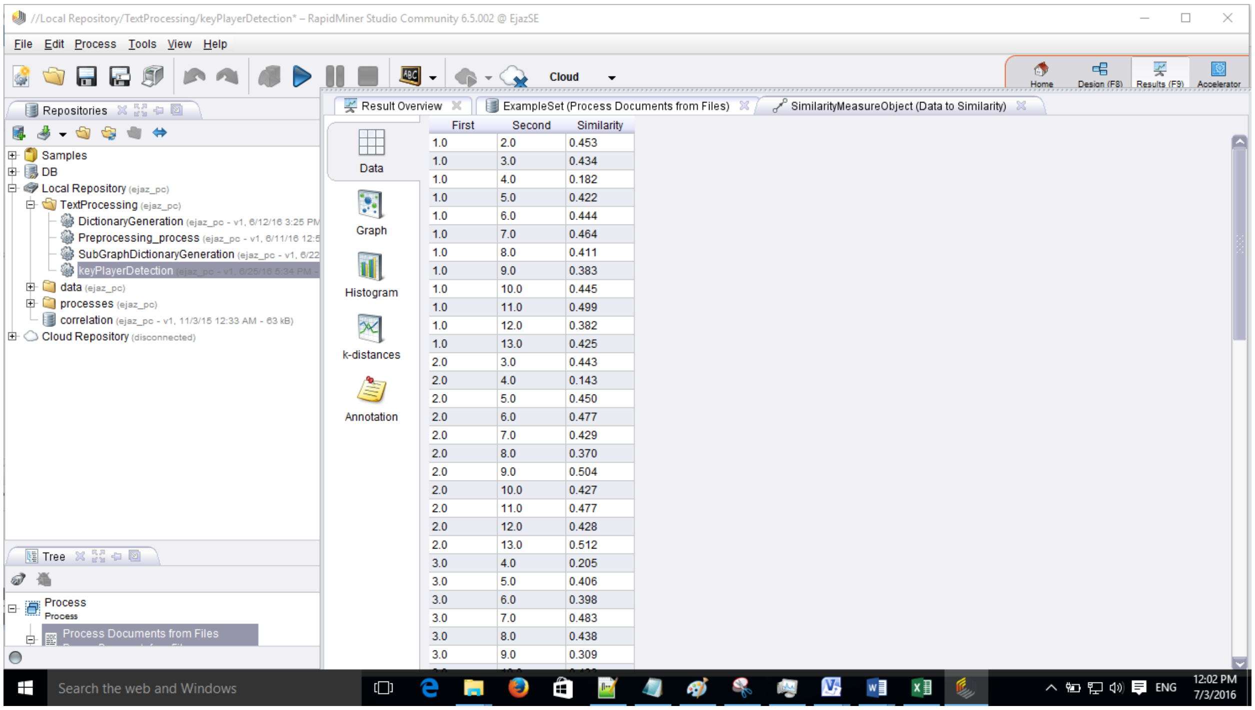
ExampleSet (13 examples, 5 special attributes, 45539 regular attributes)

Filter (13 / 13 examples): all

Row No.	id	label	metadata_file	metadata_path	metadata_d...	aaacredit	aaa
1	1	ConnectedNodes	allen_p.csv	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\alle	May 15, 201(	0	0
2	2	ConnectedNodes	arnold_j.csv	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\arn(	May 15, 201(	0	0
3	3	ConnectedNodes	arora_h.csv	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\aror	May 15, 201(	0	0
4	4	ConnectedNodes	badeer_r.cs\	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\bad	May 15, 201(	0	0
5	5	ConnectedNodes	bass_e.csv	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\bas	May 15, 201(	0	0
6	6	ConnectedNodes	beck_s.csv	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\bec	May 15, 201(	0	0
7	7	ConnectedNodes	benson_r.cs	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\ben	May 15, 201(	0	0
8	8	ConnectedNodes	causholli_m	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\cau	May 15, 201(	0	0
9	9	ConnectedNodes	Dictionary.cs	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\Dict	Jun 22, 201(	0.005	0.005
10	10	ConnectedNodes	dorland_c.cs	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\dorl	May 15, 201(	0	0
11	11	ConnectedNodes	lavorato_j.cs	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\lavo	May 30, 201(	0	0
12	12	ConnectedNodes	lay_k.csv	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\lay_	May 30, 201(	0	0
13	13	ConnectedNodes	love_p.csv	C:\Users\ejaz_pc\Desktop\EnronMailNetwork\SubGrapheInfo\love	May 30, 201(	0.014	0.014

Figure 4.3 – ExampleSet of selected nodes in rapidminer

Figure 4.4 shows the similarity scores of each node against the other node. By using the ids mentioned in figure 4.3, we detect the dictionary node and fetch the results of similarity between data dictionary and each other selected node. For this purpose we export the rapid miner results to Microsoft excel for analysis.


//Local Repository/TextProcessing/keyPlayerDetection* – RapidMiner Studio Community 6.5.002 @ EjazSE
File Edit Process Tools View Help
Cloud
Home
Design (F8)
Results (F9)
Accelerator
Repositories
Samples
DB
Local Repository (ejaz_pc)
TextProcessing (ejaz_pc)
DictionaryGeneration (ejaz_pc - v1, 6/12/16 3:25 PM
Preprocessing_process (ejaz_pc - v1, 6/11/16 12:5
SubGraphDictionaryGeneration (ejaz_pc - v1, 6/22
keyPlayerDetection (ejaz_pc - v1, 6/25/16 5:34 PM
data (ejaz_pc)
processes (ejaz_pc)
correlation (ejaz_pc - v1, 11/3/15 12:33 AM - 63 kB)
Cloud Repository (disconnected)
Tree
Process
Process
Process Documents from Files
Result Overview
ExampleSet (Process Documents from Files)
SimilarityMeasureObject (Data to Similarity)
Data
Graph
Histogram
k-distances
Annotation
Search the web and Windows
ENG
12:02 PM
7/3/2016

First	Second	Similarity
1.0	2.0	0.453
1.0	3.0	0.434
1.0	4.0	0.182
1.0	5.0	0.422
1.0	6.0	0.444
1.0	7.0	0.464
1.0	8.0	0.411
1.0	9.0	0.383
1.0	10.0	0.445
1.0	11.0	0.499
1.0	12.0	0.382
1.0	13.0	0.425
2.0	3.0	0.443
2.0	4.0	0.143
2.0	5.0	0.450
2.0	6.0	0.477
2.0	7.0	0.429
2.0	8.0	0.370
2.0	9.0	0.504
2.0	10.0	0.427
2.0	11.0	0.477
2.0	12.0	0.428
2.0	13.0	0.512
3.0	4.0	0.205
3.0	5.0	0.406
3.0	6.0	0.398
3.0	7.0	0.483
3.0	8.0	0.438
3.0	9.0	0.309

Figure 4.4 – Similarity Score between data dictionary and nodes

4.6 Correlation: Evaluation, Results and Conclusion

We tried to present results of all options used for vector creation. This will give a little bit variance because different options works in different way. So, here is another flexibility user can experience by using this model. Table 4.1 shows the similarity or correlation results between data dictionary and all selected nodes. While figure 4.5 shows the results of selected group in graphical form to a better understanding, which help us to detect the node having highest similarity score among the group. As we decided in previous chapter to choose the cosine similarity measure, which outputs similarity result between 0 and 1. We then apply percentage on the results to get the ratio of how many words are matching between data dictionary and a specific node.

In table 4.1 a row highlighted in red, which shows the **65.91%** similarity of words between data dictionary and lay-k, which is the highest among the selected group. Then, we look for the nodes which are connected to lay-k in the main network, but we find that it is only connected to allen-p. Ok, fine we did not stop here, we moved to next node having second highest similarity score which is 55.78% between data dictionary and beck-s, but unfortunately it also connected only to allen-p. We further move to next highest similarity score which is 54.96% between data dictionary and node bass-e. Now, in this case we find that there are some nodes connected with bass-e in the main network. We now find this group and apply our model on new nodes along with data dictionary.

Dictionary Label	ID	Node Label	Similarity Score	Percentage
Dictionary	1	allen_p	0.3829	38.29%
Dictionary	2	arnold_j	0.5041	50.41%
Dictionary	3	arora_h	0.3089	30.89%
Dictionary	4	badeer_r	0.0814	8.14%
Dictionary	5	bass_e	0.5496	54.96%
Dictionary	6	beck_s	0.5578	55.78%
Dictionary	7	benson_r	0.3757	37.57%
Dictionary	8	causholli_m	0.3166	31.66%
Dictionary	9	dorland_c	0.4163	41.63%
Dictionary	10	lavorato_j	0.4168	41.68%
Dictionary	11	lay_k	0.6591	65.91%
Dictionary	12	love_p	0.5229	52.29%

Table 4.1 – Similarity Score using binary term occurrences

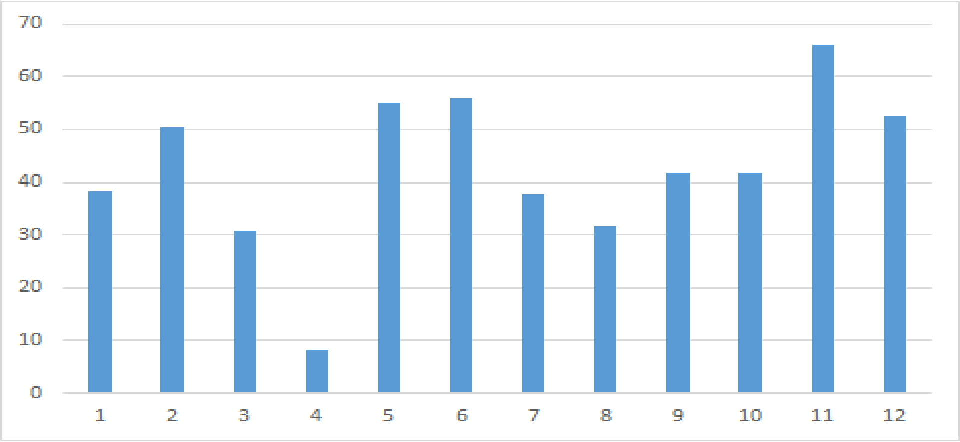


Figure 4.5 – Similarity percentage of selected group

Table 4.2 shows the results of nodes included in new group linked to bass-e. Subgroup also includes nodes like allen-p, arora-h which are both present in previous group, so we filter them out for finding their similarity with data dictionary again. For new group highlighted in below table "martin-t" has the similarity score about 47.34% which is highest, but this score is far behind the scores of some node in previous group. By following table we could conclude that no one is as key as some of the nodes in previous group.

Dictionary Label	Node Label	Similarity Score	Percentage
Dictionary	baughman-d	0.3722	37.22%
Dictionary	hodge-j	0.3072	30.72%
Dictionary	martin-t	0.4734	47.34%

Table 4.2 – Similarity Score of liked group

4.6.1 Key player detection

After subsequent test and analysis, we decided that if a node has 30% similarity score by creating word vector using Binary term occurrences or any other option, could be considered as terrorist but that does not mean it would be most key player in the node, key player would be that node which has the highest similarity score among all nodes. Top 3 nodes with highest similarity score could be the key players as per our understanding and analysis using our model. So, if we look in above tables we could say that **lay-k**, **beck-s** and **bass-e** are the key players in the network. These nodes may be responsible for every key decision and planning of the terrorist activities.

So, from our above analysis and correlation scores we could conclude and detect the key players. Table 5.3 shows the key player in the network with their similarity score with data dictionary.

Node Label	Similarity Score	Percentage
lay_k	0.6591	65.91%
beck_s	0.5578	55.78%
bass_e	0.5496	54.96%

Table 4.3 – Key players and their similarity score

4.6.2 Comparison of results

As, we discussed there are four options used to create word vectors using process documents from files. We, also analyzed the similarity score obtained by using all options. Table 4.4 shows the comparisons between all options for detecting key node. Table shows little differences between the correlation score for first three options, but difference is little big when it comes to TF-IDF. But if you look at the table and notice the highlighted in red rows, that shows the highest top three node with highest similarity score are the same for all four options and are those which we have detected as key player in above table as well. So, this indicate that whatever value is used for vector creation, it will not affect the key player detection.

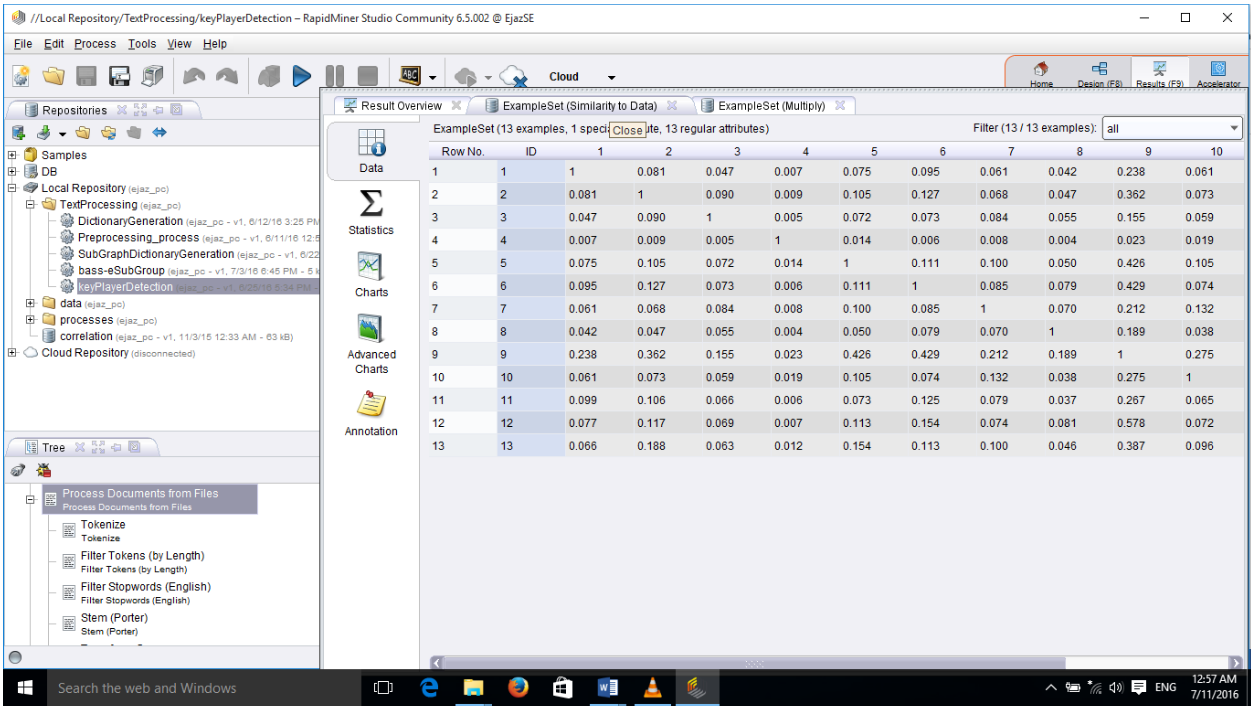
Dictionary	Node	Binary Term	Term	Term	TF-IDF
Dictionary	allen_p	38.29%	38.23%	38.23%	23.83%
Dictionary	arnold_j	50.41%	50.31%	50.31%	36.19%
Dictionary	arora_h	30.89%	30.89%	30.89%	15.52%
Dictionary	badeer_r	8.14%	8.24%	8.24%	2.29%
Dictionary	bass_e	54.96%	54.83%	54.83%	42.59%
Dictionary	beck_s	55.78%	55.61%	55.61%	42.91%
Dictionary	benson_r	37.57%	37.51%	37.51%	21.22%
Dictionary	causholli_m	31.66%	31.67%	31.67%	18.90%
Dictionary	dorland_c	41.63%	41.57%	41.57%	27.46%
Dictionary	lavorato_j	41.68%	41.65%	41.65%	26.75%
Dictionary	lay_k	65.91%	65.67%	65.67%	57.83%
Dictionary	love_p	52.29%	52.18%	52.18%	38.72%

Table 4.4 – Comparison of all vector creation options

So, that summarize the method of finding correlation between data dictionary and nodes. For the above test we have to divide our data for training and test data set to get and understand the similarity parameters and result. We used our 60% data as the training data set and 40% as test. We first applied 60% training data to our model to train it. After that we apply it on remaining 40% test set, which shows the above results.

4.6.3 Correlation among the terrorist group

We, also perform a test to get the correlation among the nodes as well. This will show the similarity between different nodes, which provides a different angle to detect the key player in network as well. However, this will only be effective when you do not have data dictionary and you are confirmed that it is a terrorist network. These analysis could help us to detect which specific nodes are communicating for same purpose if their similarity score is on higher side. As terrorist will used the specific words or code words for their planning and other actions.


//Local Repository/TextProcessing/keyPlayerDetection – RapidMiner Studio Community 6.5.002 @ EjazSE
File Edit Process Tools View Help
Cloud
Home
Design (F8)
Results (F9)
Accelerator
Repositories
Samples
DB
Local Repository (ejaz_pc)
TextProcessing (ejaz_pc)
DictionaryGeneration (ejaz_pc - v1, 6/12/16 3:25 PM
Preprocessing_process (ejaz_pc - v1, 6/11/16 12:5
SubGraphDictionaryGeneration (ejaz_pc - v1, 6/22
bass-eSubGroup (ejaz_pc - v1, 7/3/16 6:45 PM - 5 k
keyPlayerDetection (ejaz_pc - v1, 6/25/16 5:34 PM
data (ejaz_pc)
processes (ejaz_pc)
correlation (ejaz_pc - v1, 11/3/15 12:33 AM - 63 kB)
Cloud Repository (disconnected)
Tree
Process Documents from Files
Process Documents from Files
Tokenize
Tokenize
Filter Tokens (by Length)
Filter Tokens (by Length)
Filter Stopwords (English)
Filter Stopwords (English)
Stem (Porter)
Stem (Porter)
Result Overview
ExampleSet (Similarity to Data)
ExampleSet (Multiply)
Data
Statistics
Charts
Advanced Charts
Annotation
Close
Search the web and Windows
ENG
12:57 AM
7/11/2016

ExampleSet (13 examples, 1 speci Close ute, 13 regular attributes)

Filter (13 / 13 examples): all

Row No.	ID	1	2	3	4	5	6	7	8	9	10
1	1	1	0.081	0.047	0.007	0.075	0.095	0.061	0.042	0.238	0.061
2	2	0.081	1	0.090	0.009	0.105	0.127	0.068	0.047	0.362	0.073
3	3	0.047	0.090	1	0.005	0.072	0.073	0.084	0.055	0.155	0.059
4	4	0.007	0.009	0.005	1	0.014	0.006	0.008	0.004	0.023	0.019
5	5	0.075	0.105	0.072	0.014	1	0.111	0.100	0.050	0.426	0.105
6	6	0.095	0.127	0.073	0.006	0.111	1	0.085	0.079	0.429	0.074
7	7	0.061	0.068	0.084	0.008	0.100	0.085	1	0.070	0.212	0.132
8	8	0.042	0.047	0.055	0.004	0.050	0.079	0.070	1	0.189	0.038
9	9	0.238	0.362	0.155	0.023	0.426	0.429	0.212	0.189	1	0.275
10	10	0.061	0.073	0.059	0.019	0.105	0.074	0.132	0.038	0.275	1
11	11	0.099	0.106	0.066	0.006	0.073	0.125	0.079	0.037	0.267	0.065
12	12	0.077	0.117	0.069	0.007	0.113	0.154	0.074	0.081	0.578	0.072
13	13	0.066	0.188	0.063	0.012	0.154	0.113	0.100	0.046	0.387	0.096

Figure 4.6 – Similarity score matrix of selected nodes

Figure 4.6 shows the similarity or correlation score between different selected nodes in matrix form. For further analysis we export this result in excel and find the percentage of similarity which is shown in below table.
]

	allen_p	arnold_j	arora_h	badeer_r	bass_e	beck_s	benson_r	causholli_m	dorland_c	lavorato_j	lay_k
allen_p	100	45.29	43.37	18.25	42.15	44.43	46.43	41.07	44.46	49.88	38.15
arnold_j	45.29	100	44.26	14.32	44.95	47.73	42.89	37.00	42.73	47.69	42.77
arora_h	43.37	44.26	100	20.51	40.62	39.84	48.31	43.80	43.32	44.47	34.20
badeer_r	18.25	14.32	20.51	100	13.92	13.24	18.44	20.30	17.69	17.02	11.25
bass_e	42.15	44.95	40.62	13.92	100	44.81	46.64	36.08	47.39	42.36	40.70
beck_s	44.43	47.73	39.84	13.24	44.81	100	43.39	37.71	42.08	47.50	45.25
benson_r	46.43	42.89	48.31	18.44	46.64	43.39	100	44.81	54.53	47.64	36.80
causholli_m	41.07	37.00	43.80	20.30	36.08	37.71	44.81	100	39.49	39.96	34.26
dorland_c	44.46	42.73	43.32	17.69	47.39	42.08	54.53	39.49	100	44.58	36.70
lavorato_j	49.88	47.69	44.47	17.02	42.36	47.50	47.64	39.96	44.58	100	40.53
lay_k	38.15	42.77	34.20	11.25	40.70	45.25	36.80	34.26	36.70	40.53	100

Table 4.5 – Correlation between selected nodes

Above table 4.5 shows the correlation between different selected nodes. In above table highlighted shows the highest similarity scores of each node with another in table. One important point to be noted here is that similarity score between two nodes will remain same, whether you want to calculate it from node1 to node2 or node2 to node1. For example in above table node “allen-p” has its highest similarity with “lavorato-j” node. Same similarity score can be noted from “lavorato-j” to “allen-p”.

Once, we find the highest similarity scores, this could help us to detect the key players who has highest similarity score between them. For example, in above table node “dorland_c” has hightest similarity score with node “benson_r” and vice versa, and also this score is above 30% the minimum percentage threshold value for a node to lie in key player category. So, we can labeled these nodes as the key nodes as well.

Table 4.6 shows the top 3 key players detected from the table 4.5.

To	From	Similarity scorePercentage
allen_p	lavorato_j	49.88%
arora_h	benson_r	48.31%
benson_r	dorland_c	54.53%

Table 4.6 – Key player from selected nodes

Chapter 5

Conclusion and Future Work

Chapter 5

5.1 Introduction

This chapter is mainly focused on concluding the research work done in the thesis by presenting the three aspects of our work. We conclude what are our contributions in anti-terrorism field collaborating with the latest information and communication technologies, our contribution regarding research and highlighted the future consideration in the field of anti-terrorism using the communication between nodes of networks.

5.2 Conclusions

Counter terrorism has become a key area in modern age. This use of latest information and communication technologies in this area has become vital. It has transformed into an active and key area just because to save the human lives from the horrible terror attacks. There are a lot of aspects of counter terrorism and different researchers have worked in those areas and proposed different techniques and solution to counter terrorism. One of the most prominent aspect is use of social network analysis, specifically to analyze the structures of the terrorist organizations to device strategies for their destabilization, use of different machine learning and data mining practices to detect the patterns of interest that can be helpful for law enforcement agencies to destabilize such organizations. In this thesis, using some practices of data mining and social network analysis, like text mining and key player detection using nodes communication of network, some contributions are made in counter terrorism field.

A new model has been proposed “Covert network analysis to detect the key players using text mining”, to identify not only group leaders but also other players in network who belongs to terrorist. Group leaders are the one in the terrorist network, who keep their secrecy at high level by having least connected in network. Group leaders have very low values of social network analysis renowned measures like Closeness, Eigenvector, Degree and Betweenness Centrality because of having least connected. So, it was difficult to find out group leaders using these measures. But these group leaders are connected directly having a strong relation with the people who have high values for above mentioned SNA measures. Despite the fact group leaders having least connected, they do communicate with particular individuals in some defined patterns or words. Using text mining to find correlation or similarity between a pre-defined terrorist dictionary and the communication of individuals in SNA those group leaders having low SNA measures and other terrorist can be detected easily. The proposed methodology has been validated on Enron email network.

Another novel model was proposed in the field of counter terrorism for key player detection in SNA of terrorist networks. As mentioned in literature review, all the customary measures of social network analysis are used to find out the key players pondering on different aspects of network. But in this research a hybrid model which group different text mining and documents mining techniques, is proposed to detect key players rather than the traditional measures of SNA. The proposed model is implanted and validate on Enron email dataset. The detailed analysis was performed which include quantitative as well as comparative.

By using all the work and facts and figures available a pre-defined dictionary was prepared to detect the key players. Another very important achievement in this thesis is that different key players can be grouped into different clusters based on the cosine similarity measure. This can help to detect the most influential terrorist at one place rather than finding them one by one.

5.3 Contributions

For this novel approach using nodes communication, some contribution made to accomplish the task. These incudes preparation of data dictionary which used as a reference point to find the key players using their communication, a common text mining module which will be used for transforming any unstructured data into structured data, labeling and constructing Enron email network and a new novel model to detect key players (terrorist leader) in social network using data dictionary, text mining module and different rapid miner operators.

This model could benefit law and enforcement agencies to analyze social network using their contents and a reference data dictionary which is the highly important ingredient of this research and for the success of the model.

5.4 Future Work

Future enhancements are the functionalities that can be further added to system to increase its performance and functionality domain.

Chapter 5

One of the major future enhancements for the model is to detect what they are communicating about, their planning, their future activities etc, using their communication apart from just detecting group leader to destabilize the network. Also this could lead us to detect the major actors involved in the network, another potential future enhancement. An automated result analysis module would be more helpful. This could be another future enhancement.

This model could be used to devise another novel model by integrating this model with any other social network analysis method which used centrality measure to detect key players for more better results and efficiency as this field is becoming more and more important for the safety and security of masses. So, it would be much better if we get the much and accurate results.

References

REFERENCES

[1] Wasi Haider Butt, M. Usman Akram, Shoab A. Khan, andMuhammad Younus Javed, “Covert Network Analysis for Key Player Detection and Event Prediction Using a Hybrid Classifier”, Hindawi Publishing Corporation, 2014

[2] Louise Lemieux-Charles, Larry W. Chambers, Rhonda Cockerill, Susan Jaglal, Kevin Brazil, Carole Cohen, Ken LeClair, Bill Dalziel and Barbara Schulman, “Evaluating the Effectiveness of Community-Based Dementia Care Network s: The Dementia Care Networks’ Study”, 2005.

[3] Mahyuddin K. M. Nasution1 and Maria Elfida1, “Terrorist Network: Towards An Analysis”, PS Sistem Informasi, STTH, Medan, Sumatera Utara, Indonesia, 2011

[4] V. E. Krebs, "Mapping Networks of Terrorist Cells," INSNA, vol. 24, no. 3, pp. 43-52, 2002.

[5] "http://www.orgnet.com/sna.html," [Online]. Available: http://www.orgnet.com/sna.html. [Accessed April 2013].

[6] L. C. Freeman, " The Study of Social Networks," [Online]. Available: http://www.insna.org/INSNA/na_inf.html. [Accessed April 2013] .

[7] A. H. Dekker, "Centrality In Social Networks: Theoretical And Simulation Approaches," SimTecT, 2008.

[8] M. B. R. P.-S. a. A. V. A. Barrat, "The architecture of complex weighted networks," National Academy of Sciences, vol. 101, no. 11, p. 3747–3752, 2004.

[9] F. A. ,. S. Tore Opsahl, "Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths," Social Networks, 2010.

[10] Ahmet Erdem Sarıyüce, Kamer Kaya1, Erik Saule1 and Umit V. Catalyurek1, “Incremental Algorithms for Network Management and Analysis based on Closeness Centrality”, arXiv:1303.0422v1 [cs.DS], 2013.

[11] P. B. a. P. Lloyd, "Eigenvector-like measures of centrality for asymmetric relations," Social Networks, vol. 23, pp. 191-201, 2001.

[12] A. Dekker, "Conceptual Distance in Social Network Analysis," Journal of Social Structure, vol. 6, no. 3, 2005.

[13] P. &. H. F. Hage, "Eccentricity and centrality in networks," Social Networks, vol. 17, p. 57–63, 1995.

References

[14] L. B. S. &. W. D. Freeman, "Centrality in valued graphs: A measure of betweenness based on network flow," Social Networks, vol. 13, p. 141–154, 1991.

[15] G. Barbian, "Trust Centrality in Online Social Networks," in European Intelligence and Security Informatics Conference, 2011.

[16] N. M. a. H. L. Larsen, "Investigative Data Mining Toolkit: A Software Prototype for Visualizing, Analyzing and Destabilizing Terrorist Networks," in Visualising Network Information (pp. 14-1 – 14-24). Meeting Proceedings. RTO-MP-IST-063, Paper 14. Neuilly-sur-Seine, France, 2006.

[17] J. S. a. J. Adibi, "Discovering Important Nodes through Graph Entropy, The Case of Enron Email Database," in Illinois ACM, Chicago, 2005.

[18] S. Karthika, "Identifying Key Players in a Covert Network using," in IEEE International Conference on Recent Trends in Information Technology, 2012.

[19] D. B. Skillicorn, "Social Network Analysis via Matrix Decompositions : al Qaeda," School of Computing Queen's University, 2004.

[20] Tanu Verma, Renu and Deepti Gaur, "Tokenization and Filtering Process in RapidMiner", International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 7– No. 2, 2014.

[21] Anna Huang, "Similarity Measures for Text Document Clustering", NZCSRSC 2008, Christchurch, New Zealand, 2008.

[22] D. Gomez, "Centrality and power in social networks: a game theoretic approach," Mathematical Social Sciences, vol. 46, pp. 27-54, 2003.

[23] R. L. a. I. Blankers, "Key player identification : a note on weighted," in International Conference on Advances in Social Networks Analysis and Mining, 2010.

[24] K. M. Carley, "Destabilization of Covert Networks," Computational & Mathematical Organization Theory, vol. 12, no. 1, pp. 51-66, 2006.

[25] N. R. S. Y.Narahari, "Determining the Top-k Nodes in Social Networks using the Shapley Value," in Int.Conf.on Autonomous Agents and Multiagent Systems , Portugal, 2008.

[26] J. Farley, "Breaking Al Qaeda Cells: A Mathematical Analysis of Counterterrorism Operations (A Guide for Risk Assessment and Decision Making)," Studies in Conflict and Terrorism, vol. 26, pp. 399-411, 2003.

References

[27] J. R. N. K. Kathleen M. Carley, "Destabilizing Terrorist Networks," in NAACSOS, Pittsburgh, 2003.

[28] "http://www.elearningpost.com/articles/archives/qa_with_professor_karen_stephenson/," [Online].Available: http://www.elearningpost.com/articles/archives/qa_with_professor_karen_stephenson/. [Accessed May 2013].

[29] Ms.K.Sruthi and Mr.B.Venkateshwar Reddy, "Document Clustering on Various Similarity Measures" , International Journal of Advanced Research in Computer Science and Software Engineering, 2013.

[30] Pontus Svenson, Per Svensson andHugo Tullberg, "Social network analysis and information fusion for anti-terrorism", CIMI2006, 2006.

[31] S.Karthika, A.Kiruthiga and S.Bose, "Dominant Features Identification for Covert Nodes in 9/11 attack using their Profile", International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.4, July 2012.

[32] Kaltrina Nuredini, Ozcan Asilkan and Atixhe Ismaili, "An Exemplary Survey Implementation on Text Mining with Rapid Miner", 1st International Symposium on Computing in Informatics and Mathematics (ISCIM 2011), 2011.

[33] Marc A. Smith, Ben Shneiderman, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer and Eric Gleave, "Analyzing (Social Media) Networks with NodeXL", C&T'09, 2009.

[34] Clifford Weinstein, William Campbell, Brian Delaney and Gerald O'Leary, "Modeling and Detection Techniques for Counter-Terror Social Network Analysis and Intent Recognition", IEEE, 2009.